IP LodB project

FINAL REPORT to the EPO

with attached main Deliverables

IP LodB

2020

Contents

This report has been prepared by the members of the IPLOD group: Dolores Modic, Alan Johnson, Borut Lužar, Ana Hafner, Borut Rožac, Miha Vučkovič and Einar Rasmussen for the European Patent Office (EPO) as part of the reporting inside the IP linked open data: building bridges (IP LodB) project. The IP LodB project is an EPO Academic Research Programme project (https://www.epo.org/learning/materials/academic-research-programme.html).

Please cite as: Modic, D., Johnson, A. Lužar, B., Hafner A., Rožac B., Vučkovič M. And Rasmussen E. (2020). IP LodB project
Final report for EPO Academic Research Programme. Available at: https://www.epo.org/learning/materials/academic-research-programme/research-project-grants.html.

The citations for the attachments are provided in the appendices.

ENHANCING THE LINKED OPEN DATA (LOD) SPACE

The IP LodB Project:

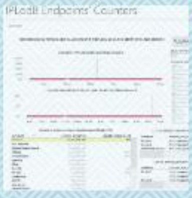
Discover more (interlinked and consolidated) patent related data in a new format

Stepping towards
Web of data
Link as a bearer of meaning
Inventor-->Author of Publications--> Article

The IP related LOD landscape revealed

What's out there? The IP LodB Map

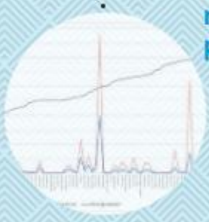
- Several database types (patents, publications etc.) building an EPO LOD database-centered sub-cloud with additional information available
- Easy to discover: The map for easy visualization of available LOD database offerings: https://www.iplod.io/Home/Map



How accessible is LOD? Databases' diagnostics

- Pinging databases to reveal their accessibility, the changes in their size and data shemas
- Available at: https://www.iplod.io/Home/MapCount

Better data: Diligently aligning LOD databases



Increasing uniqueness of real-life objects (i.e. persons) in LOD

- Advanced disambiguation
- Reduction in unique persons
- Using the iplod nascent ID identifier in other EPO data products for natural persons (inventors, applicants etc...)

Interconnecting authors & inventors

- Aligning inventors' records from EPO LOD to authors' records from Springer Nature LOD
- Providing IPLOD Entity cards with consolidated data on inventorship and publication activity

Discover more: exploiting interconnected data

...wse IP LodB Un...

...rch by **family name** or **IPLodB ID**:

	Full Name (one of)
●	Kostelnik, Callen E.
●	Kostelnik, Dr. Jan
●	Kostelnik, Jan
	Kostelnik...

More data discovery!

- Data connected to further LOD databases; e.g. GRID data on affiliations
- An available data search query and bulk download (RDF links)
- At: https://www.iplod.io/Persons

Cleaner data for scientific research & practice

- Gather information on and understand inventive and publishing data and their interconnections
- Connect data to test theory
- Add your own data = join us or enrich own LOD data!

IP linked open data: building bridges (IP LodB) project was co-financed by the European Patent Office Academic Research Programme (EPO ARP) in the period of 2019-2020.

Contact us at: admin@iplod.io
The team: Dolores Modic, Alan Johnson, Borut Lužar, Ana Hafner, Borut Rožac, Miha Vučkovič and Einar Rasmussen

IPLOD
www.iplod.io

EXECUTIVE SUMMARY

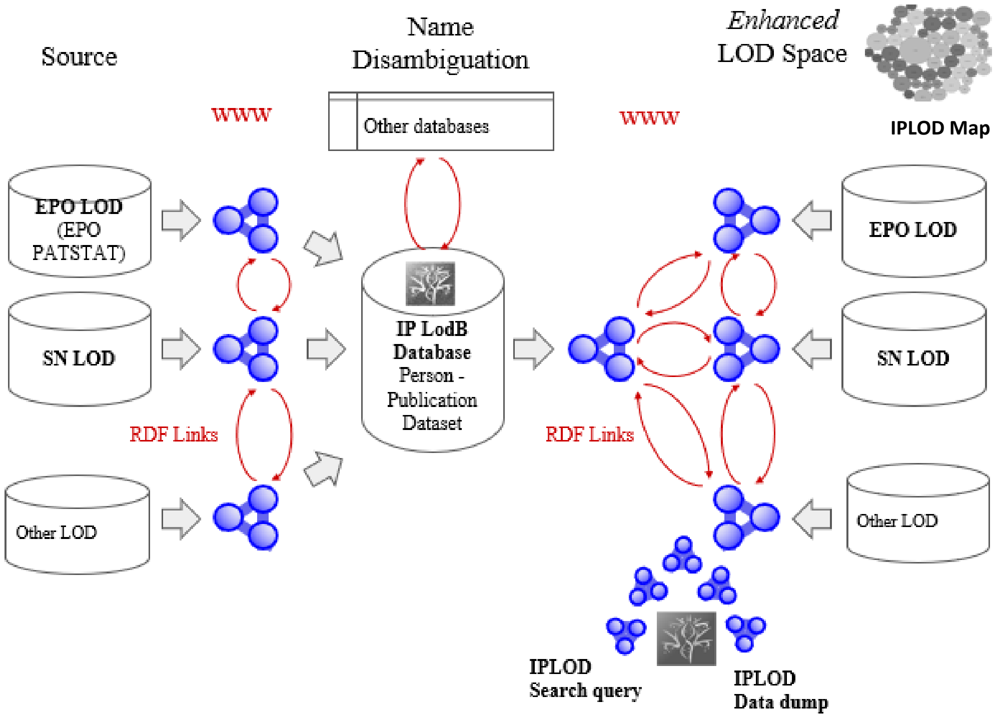
IP **LodB capitalized on the linked open data (LOD) thinking**, to build bridges between patented information and scientific knowledge, whilst **focusing on individuals** who codify new knowledge and their connected organizations, including those who apply patents in new products and services. The project funding allowed for the formation of **an interdisciplinary team**, with researchers from the field of data science and social sciences (both with interest in IP and in methods) to further enhancing the LOD space.

What is LOD about?

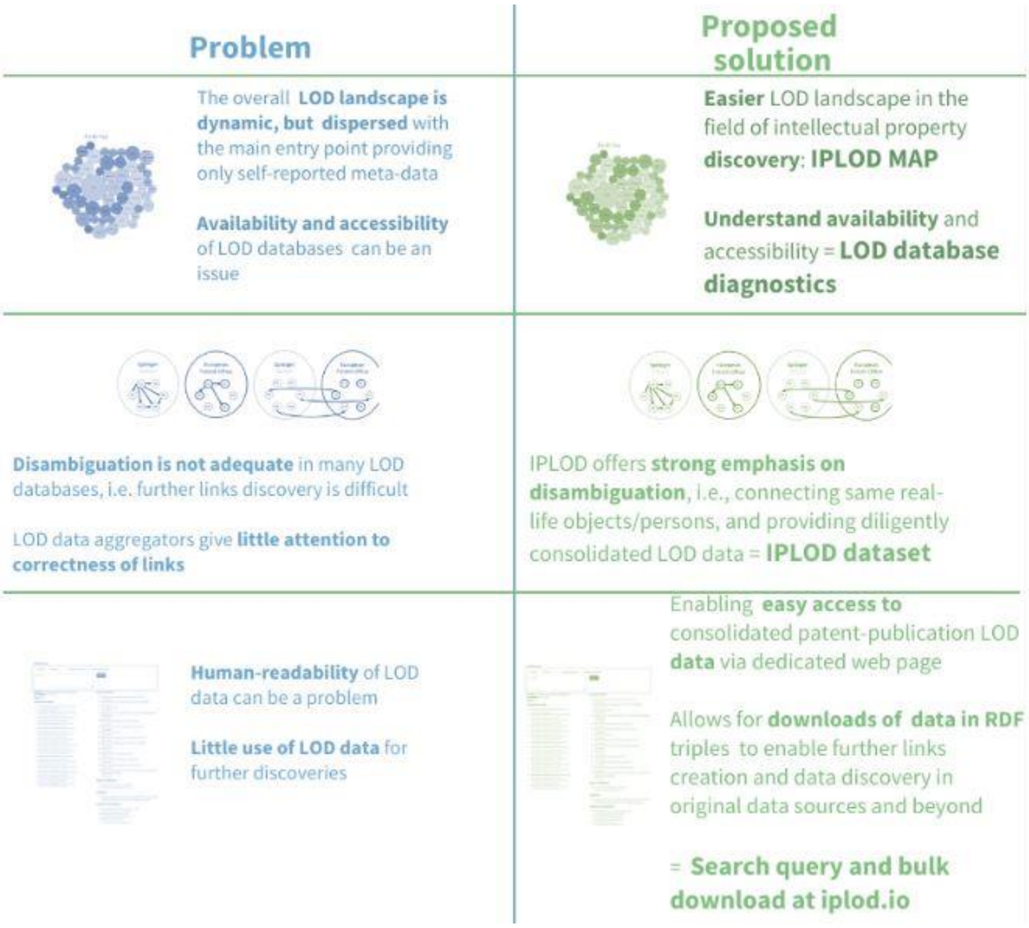
LOD protocols make data publicly available and if interlinked they can address the problem of integrating heterogeneous data, which is an inherent problem of big data (Sleeman et al, 2015). The idea is that same real-life object records would be connected, thus providing mora information discovery.

The final vision for the technology is that all the data available on the Web will be treated and researched as one database with the aim to share and reuse existing data (Khusro, 2014) while propagating a machine-readable format.

As main outputs the IP LodB produced an **intellectual property rights (IPR) linked open data (LOD) map (IP LodB Map)**, and has tested the linkability of the European patent (EP) LOD database with primarily Springer Nature, whilst increasing the uniqueness of data using different harmonization techniques, which led to the construction of the **IPLOD database/dataset.**



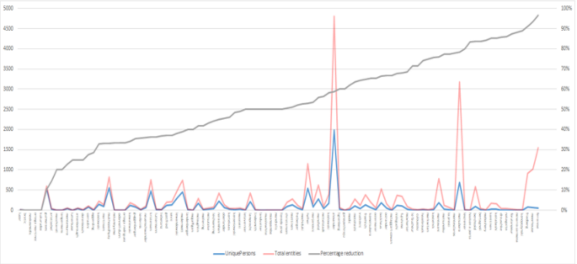
The linked open data space, albeit dynamic still suffers from several shortcomings, which impede its uptake. The IP LodB project aimed at surpassing some of the identified shortcomings.



Problem	Proposed solution
The overall **LOD landscape is dynamic, but dispersed** with the main entry point providing only self-reported meta-data **Availability and accessibility** of LOD databases can be an issue	**Easier** LOD landscape in the field of intellectual property **discovery: IPLOD MAP** **Understand availability** and accessibility = **LOD database diagnostics**
Disambiguation is not adequate in many LOD databases, i.e. further links discovery is difficult LOD data aggregators give **little attention to correctness of links**	IPLOD offers **strong emphasis on disambiguation**, i.e., connecting same real-life objects/persons, and providing diligently consolidated LOD data = **IPLOD dataset**
Human-readability of LOD data can be a problem **Little use of LOD data** for further discoveries	Enabling **easy access to** consolidated patent-publication LOD **data** via dedicated web page Allows for **downloads of data in RDF** triples to enable further links creation and data discovery in original data sources and beyond = **Search query and bulk download at iplod.io**

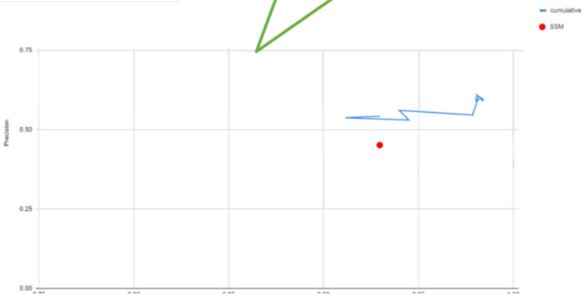
The Map deliverable has guided the researchers towards inclusion of datasets, and the understanding of the availability and *de facto* usability of various datasets, whilst additionally providing .

The IP LodB project focuses furthermore on interlinking the EPO LOD and SN LOD database, whereas it also includes some other identified LOD databases (e.g. GRID, Crossref) to build further links in the LOD space. Other databases, not currently in the LOD format, are also used to enrich the IP LodB Space to begin building an LOD innovation sub-cloud around the EPO LOD database as the hub. At the same time linkeability across the LOD innovation data space increases and the results can cross-fertilize the existing LOD datasets, if adopting the suggested links. **The IPLOD dataset** provides **consolidated, interlinked and interlinkable data.**



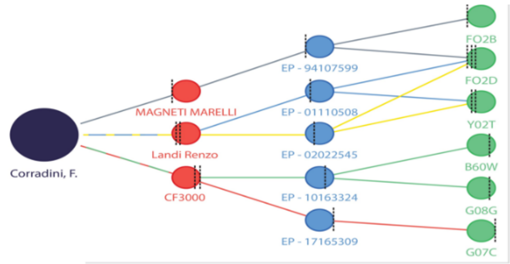
Reduction in unique persons pre- and post-disambiguation (n=25958)

Interconnecting EPO LOD and SN LOD at the rate of approx. 1.9 %

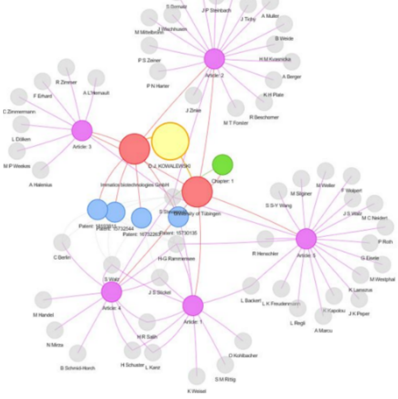


Precision and recall – comparison to simple string match (SSM)

The data is easily added to, which is the one of the benefits of the linked open data approach. This also allows the original producers of the data to be in charge of the data production, and have access to the disambiguation and interconnection results, which they can adopt. On the other hand, solutions such as ours provide for diligent IP LodB disambiguation and interconnection, which furthermore allows for various further analysis to be done.



Similar as other patent data: Inventive activity through different attributes and time

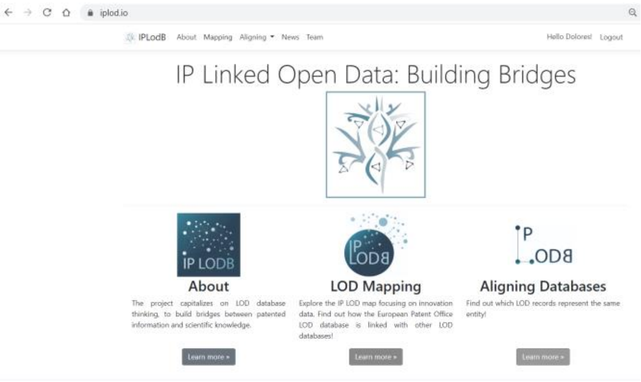


Value of linked data: Inventive and publishing activity – a network perspective

There is a need to understand the connections and data sources: the prize is also to allow LOD sources to be more accessible for theory driven research

Application on a selected field: Circular Economy (CE)

The results of the IPLOD project are fed back to the Wide World Web, both via an operational **queryable search option**, and later **through data dumps**. This is an iterative approach, so the data will become more and more complete as our work progresses. The IP LodB map and the interface to the IP LodB database are **openly available** at [iplod.io,](iplod.io) for registered and non-registered users.



IP Linked Open Data: Building Bridges

About

The project capitalizes on LOD database thinking, to build bridges between patented information and scientific knowledge.

Learn more »

LOD Mapping

Explore the IP LOD map focusing on innovation data. Find out how the European Patent Office LOD database is linked with other LOD databases!

Learn more »

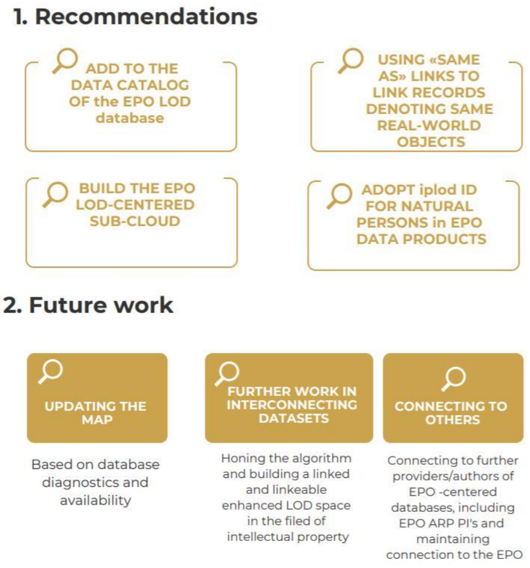
Aligning Databases

Find out which LOD records represent the same entity!

Learn more »

The knowledge was disseminated on various occasions and in different formats: as presentations on conferences (both IP oriented, e.g. EPIP) and data science oriented (e.g. ITIS), as well as workshops or parts of workshops on events meant for SMEs and other users (e.g. NIPO workshop and the KnowING IPR event Patenting and IPR management). We also expect future workshop presentations for users (consult the iplod.io webpage for further information).

Inside the methodology reports attached to this final report we provide a series of recommendations and ideas for future work to further enhance the LOD space in the field of intellectual property, and achieve more traction for this dynamic, but under-developed, LOD environment to grow. We believe one of our contributions is also bringing the disambiguation to the forefront of the LOD debates.



1. Recommendations

ADD TO THE DATA CATALOG OF the EPO LOD database

USING «SAME AS» LINKS TO LINK RECORDS DENOTING SAME REAL-WORLD OBJECTS

BUILD THE EPO LOD-CENTERED SUB-CLOUD

ADOPT iplod ID FOR NATURAL PERSONS in EPO DATA PRODUCTS

2. Future work

UPDATING THE MAP

Based on database diagnostics and availability

FURTHER WORK IN INTERCONNECTING DATASETS

Honing the algorithm and building a linked and linkeable enhanced LOD space in the filed of intellectual property

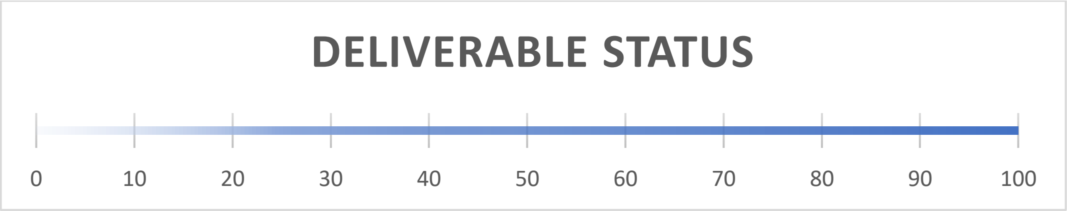
CONNECTING TO OTHERS

Connecting to further providers/authors of EPO -centered databases, including EPO ARP PI's and maintaining connection to the EPO

The final report proceeds as follows. First part is a brief report on the project management and reports on the deliverables connected therewith. The second part provides an overview of our Map related efforts, with the details available from the document 2.1. The third part is providing an overview of deliverables connected to the IPLOD database, with details available from separate document on methodology report (deliverable 3.1) and connected results (deliverable 3.2). The fourth part is connected to reporting on the dissemination deliverables. Attachment 1 provides the

PART 1: IP LOD MAP

Deliverable 2.1: Constructing the IP LOD map



The Deliverable is described in a separate document (2.1), which reports on the methodology and the result. We provide a general **EP LOD centered map** (the potential of EP LOD as the hub), the visualization of the links and the datasets availability checks. We visualize the MAP with the combination of Tableau and Rshiny. We also investigated briefly the SPARQL opportunities.

As an answer to problems of availability of the LOD databases, we also provide an assessments of selected databases' availability, i.e. LOD **database diagnostics** (SPARQL endpoint pinging)

The report on the Deliverable 2.1 is included as part of this report.

HIGHLIGHTS - MAP:

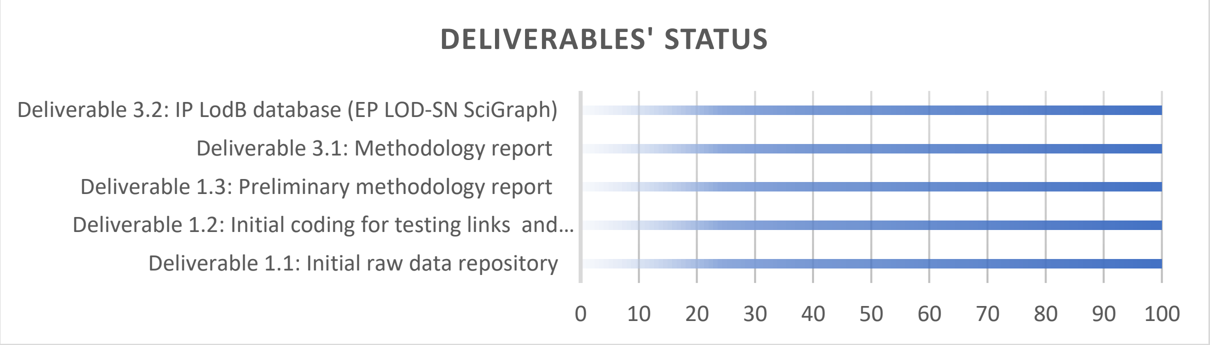
The MAP output is available at: *https://www.iplod.io/Home/Map*

We provide LOD database diagnostics on the iplod.io webpage.

The map with additional datasets represents the blueprint to continue the work on designing and providing an EP LOD-centric dataset connecting the EP LOD data with other innovation-related data.

The SPARQL opportunities remain low at this point, for various reasons, including those connected to timeouts. Nonetheless, we include(d) them in the external workshops with potential end users.

PART 2: Linking EP LOD with SN SciGraph LOD



Deliverables 1.1 - 1.3

Starting with entity disambiguation the data contains records of unique persons and their relations to existing entities in EP LOD and Springer Nature (SN) SciGraph. For faster manipulation and computation of auxiliary variables, we store those also locally. Currently, we also use other LOD databases, e.g. GRID, to output additional information about person affiliations, which is available via integrated links. We also incorporate an “on the fly” search and information from the Crossref database. Furthermore, to aid the dissemination efforts some additional variables from PATSTAT are being used (e.g. address), but available only to registered users.

The preliminary intra-database check has identified the switch in the format of the SN Graph to json format, potential auxiliary variables, the basic ontology map which can be used for the in between database alignment and the initial calculations on missing attributes. The preliminary methodology report was drafted and delivered on time to the EPO. The preliminary methodology report included both information on our IP LOD map work as well as the IPLOD database. Additional datasets and preprocessing needs were identified and several preliminary tests run with crude methodology. More on data sources, preprocessing, alignment and matching is available from the **3.1 Methodology report, which is attached to this final report.**

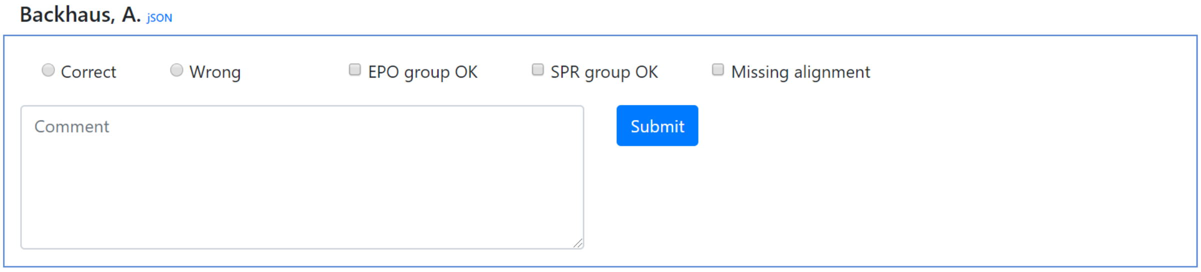
Deliverables 3.1 & 3.2

A separate document is delivered describing the IPLOD approach and results (Deliverable 3.1, the Methodology report). The disambiguation and data fusion (T3.2 and T3.3) are especially well documented, with the protocol (T3.1) as well as the validation are described (T3.2). Please note the IPLodB team has adopted in this final methodology report the terminology which is more common in disambiguation literature. This has been done especially due to own publication plans, which are based to a large degree on the methodology report.

The IPLodb database interface to the database is available at: **https://www.iplod.io/Persons.**

To incorporate some of the suggestions from the EPO workshop and include the possibility of crowdsourcing a registration is available upon request (admin controlled) which allows annotations to the suggestions. Cookies policy needs to be accepted beforehand.

Figure: Annotation for training set (i.e. golden dataset)



Backhaus, A. JSON

○ Correct ○ Wrong □ EPO group OK □ SPR group OK □ Missing alignment

HIGHLIGHTS:

The user interface with the search query is available at: https://www.iplod.io/Home/Map - crowdsourcing is enabled via annotation possibility for registered users.

The IPLOD approach is described in the methodology report, with special emphasis on disambiguation and alignment efforts.

A hand curated golden dataset (training sample) is created with almost three thousand hand annotated EPO and SciGraph original (i.e. pre-disambiguation) persons, which allowed for validation, and constitutes the training set for machine learning.

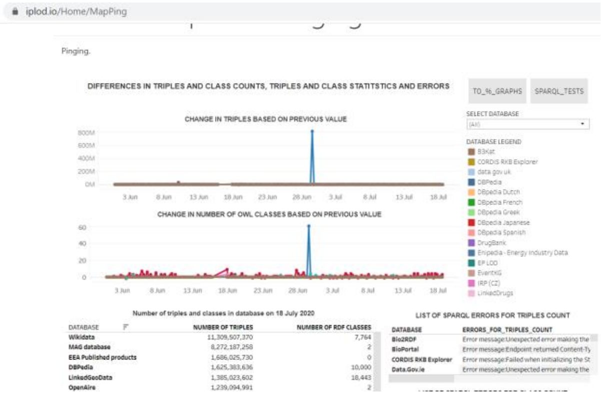
PART 3: Dissemination, Project management, quality control and reporting

Deliverables 5.1 – 5.7: Dissemination

The main dissemination channel is the dedicated webpage available at http://www.iplod.io/. The user interface allows accessing the IPLodB database, as well as there is a link for bulk download (RDF links for version 110 sample is available – see for more also Deliverable 3.1).

The webpage offers a queryable interface to the openly available IPLOD data and is being continuously enriched. The registered users are also able to query through previous versions. For registration users need to submit the registration and we suggest also writing to admin@iplodb.io.

Secondly, the map and related webpage diagnostics is available on the iplod.io webpage under Mapping (Mapping→ IPLodB Map; Mapping→ Endpoint Counts; Mapping→ Endpoint Pings).



Pinging.

DIFFERENCES IN TRIPLES AND CLASS COUNTS, TRIPLES AND CLASS STATITSTICS AND ERRORS

Number of triples and classes in database on 18 July 2020

DATABASE	NUMBER OF TRIPLES	NUMBER OF RDF CLASSES
Wikidata	11,309,507,370	7,764
MAG database	8,272,187,258	2
EEA Published products	1,686,025,730	0
DBPedia	1,625,383,636	10,000
LinkedGeoData	1,385,023,602	18,443
OpenAire	1,239,094,991	2

LIST OF SPARQL ERRORS FOR TRIPLES COUNT

DATABASE	ERRORS_FOR_TRIPLES_COUNT
Bio2RDF	Error message:Unexpected error making the
BioPortal	Error message:Endpoint returned Content-Ty
CORDIS RKB Explorer	Error message:Failed when initializing the St
Data.Gov.ie	Error message:Unexpected error making the

The complete IPLOD rdf linkages bulk download will be available after the publication stream. Currently a sample version for version 130 is available on the webpage at https://www.iplod.io/Home/Download. Upon completion, the database will be reported to the LOD cloud at https://lod-cloud.net/.

IPLODB team engaged in various dissemination efforts, whereas the publication of results in repositories is both in connection to releasing the results on our webpage (initial batch is available for testing purposes) as well as publishing the tools on repositories. The series of PPT on linked open data are available both at the iplod.io webpage (News section) and on SlideShare (https://www.slideshare.net/IPLODProject/linked-open-data-lod-part-1, https://www.slideshare.net/IPLODProject/linked-open-data-lod-part-2, https://www.slideshare.net/IPLODProject/linked-open-data-lod-part-3). The list of main events where the project has been presented is available at https://www.iplod.io/Home/News. Further publication efforts are pending.

The IPR LOD Handbook is openly published on the iplod.io webpage (under About→ Handbook). The EPIP conference and the International Conference on Information Technologies and Information (ITIS) conference were attended in 2019.

Together with the Norwegian Industrial Property Office (NIPO) the IP LodB team held a workshop in the field of IP LOD in Oslo, Norway on February 11, 2020. This was a half-day workshop. Link to event: https://www.facebook.com/events/526128264636616/permalink/534174900498619/

The next LOD workshop is scheduled for May 2021; more information on forthcoming workshops will be published on the iplod.io webpage (news section).

HIGHLIGHTS DISSEMINATION:

The workshop for users was conducted in cooperation with the Norwegian Industrial Property Office for a variety of users.

Conferences and events were attended with presentations on the project.

Several open source contributions; from handbook, to tool, PPTs etc.

Deliverable 4.1 – 4.4: Project management

There are 4 types of team meetings: IPLODB internal workshops, IPLODB team meetings, MAP team meetings and database alignment meetings. Committee meetings also involve Prof. Einar Rasmussen (Nord University) for quality control. All minutes for committee meetings, IPLODB team meetings and IPLODB workshops are available upon request. All relevant contracts have been signed at the beginning of the project.

IPLODB workshops were designed to assure common understanding of goals and tasks and to advance joint work. The workshops were done in person and on-line. In end February the workshop was cancelled due to coronavirus crisis. All workshop at the end of the project are conducted online due to travel restrictions. The table of workshops is available in Appendix 1.

The team also gratefully acknowledges the cooperation of M. Vučkovič (joining the work on map and disambiguation in accordance to his programming skills) and S. Modic (providing in-kind logos and other graphical documents for the webpage). We also acknowledge the generous support by NORD University Business School for covering all necessary travel and accommodation costs related to workshops.

Two reports were drafted, the Intermediate report (delivered timely to the EPO) and this Final report. The first workshop in December of 2019 was attended with the interim project presentation; and the second in January of 2021 with the final project presentation.

HIGHLIGHTS – PROJECT MANAGEMENT, TEAM WORK:

The team has connected to other groups inside the EPO ARP projects and to the EPO staff.

The team remains committed to continue collaboration and has already invested efforts to continue advancing the IPLOD effort also by further project proposals.

The team was able to connect the work on IPLOD with other research interests and plans to use the IPLOD data to continue their research.

Appendix 1

IP LodB project

IP LOD MAP

(Deliverable 2.1)

2020

This report has been prepared by the members of the IPLOD group: Dolores Modic, Borut Rožac, Ana Hafner, Alan Johnson, Borut Lužar, Miha Vučkovič and Einar Rasmussen for the European Patent Office (EPO) as part of the reporting inside the IP linked open data: building bridges (IP LodB) project. The IP LodB project is an EPO Academic Research Programme project (https://www.epo.org/learning/materials/academic-research-programme.html).

Please cite as: Modic, D., Rožac B., Hafner A., Johnson, A., Lužar, B., , Vučkovič M. and Rasmussen E. (2020). IP LodB project IP LOD Map. Report for EPO Academic Research Programme. Available at: https://www.epo.org/learning/materials/academic-research-programme/research-project-grants.html.

Executive Summary

In principle, LOD protocols make data publicly available and linked to known databases, as well as it can address the problem of integrating heterogeneous data, which is an inherent problem of big data. The final vision for the technology is that all the data available on the Web will be treated and researched as one database with the aim to share and reuse existing data through a machine-readable format.

While the amount of Linked Open Data has grown significantly in recent years, most data sources are still not sufficiently interlinked. Difficulties in identifying entities in same or different data sources that describe the same real-world object persist, many are difficult to discover and their availability/accessibility are unstable. Most challenges stem from the nature of LOD: it is open and distributed, but on the other hand it is plagued by low connectivity. Its ontology-based approach allows machines to understand data, however the tools allowing humans to browse through the data are lacking. The technical structure makes it demanding even to search through SPARQL which is standardized query language for querying LOD data, but compared to SQL query language which is oriented to querying tabular data, SPARQL is oriented to querying graph patterns of LOD data thus requiring good knowledge of LOD dataset structure. There is also a lack of proper development tools for querying data with SPARQL, libraries for mainstream programming languages, and database engines for storing LOD data. One of the biggest disadvantages compared to SQL is much smaller community of users and developers thus preventing it from becoming more popular.

The **IP LOD Map** focuses on the IP and related innovation data, bringing additional information on their attributes. The IP LOD Map per se brings a first EPO-centric LOD map, and can serve as a guidance for those interested in innovation data in LOD format. Inside the IP LodB project the biggest value of the IP LOD map is as a supporting deliverable, which supports our bridging approach (see for more IP LodB Methodology report). It influenced how we proceeded with aligning (i.e. bridging) the two focal databases) as well as decisions on further linkages in the alignment process. We also provide some ideas on the general interlinking potential. This deliverable provides protocols used for both. The Map is being continuously updated, with databases being retracted (e.g. if not active enough) or added, if new relevant databases come to existence.

To provide more information on the accessibility and availability of the relevant databases a pinging protocol has been developed, i.e. the **LOD database diagnostics**. Not only are the diagnostics helpful in determining the accessibility, they also provide additional information on the size, growth and (re-)classifications.

We recommend further development of IP LOD Map with providing information on further relevant databases to make this EP LOD centric LOD map more complete. Also, to encourage researchers to provide the datasets in compatible formats, so they can be merged to further enhance the LOD space. We welcome all further suggestions: please write to: dolores.modic@nord.no or admin@iplod.io.

Introduction to Linked Open Data and its promises

The semantic web emphasizes a move away from publishing documents readable by humans and towards publishing data readable by machines. The general idea was provided by Tim Bernes-Lee, one of the "fathers" of the world wide web (WWW). To be able to talk about the linked (open) data the following four rules need to be observed (Berners-Lee, 2006) :

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs so that they can discover more things.

In principle, LOD protocols make data publicly available and linked to known databases, as well as it can address the problem of integrating heterogeneous data (Sleeman et al., 2013). The final vision for the technology is that all the data available on the Web will be treated and researched as one database with the aim to share and reuse existing data (Khusro et al, 2014) while propagating a machine-readable format. However, while the amount of Linked Open Data has grown significantly over the last years**, most data sources are still not sufficiently interlinked, the difficulties in identifying entities in same or different data sources that describe the same real-world object persist, many are difficult to discover and their availability/accessibility is unstable.**

Most challenges stem from the nature of LOD: it is open and distributed, but on the other hand it is plagued by low connectivity. Its ontology-based approach allows machines to understand data, however the tools allowing humans to browse through the data are lacking. The technical structure makes it demanding even to search through (SPARQL being the language to query data).

From LOD cloud to particular LOD maps

Firstly, we are faced with the problem of discoverability. As Assaf et al (2006) have written considering the variation in size, languages and the actualizations of the data, finding useful datasets without prior knowledge is increasingly complicated.

The Linked Open Data Cloud[1] shows 1,239 datasets published (March 2019), from 93 datasets ten years ago, with many reputable institutions, e.g., European Patent Office (EPO), adopting the linked open data (LOD) format to publish data. LOD data sources are potentially powerful and cost-effective business intelligence resources for new and small businesses. The Linked Open Data Cloud provides the metadata and albeit often criticized (e.g. Polleres et al, 2020)

Within this cloud of LOD databases, DBPedia serves as a "nucleus" of LOD (Auer et al, 2007) with more than 28 billion unique triples in 2017 (Fernandez et al., 2017) and sets RDF links pointing into various external data sources. Thus, many other LOD publishers decided to set RDF links pointing to DBpedia from their data sets; making the DBpedia a central inter-linking hub in the Web of Linked Data. Some authors claim DBPedia was key to success of the LOD initiative (Lehmann et al, 2012). On the upside, DBpedia supplies authoritative references for ubiquitous data dimensions (Auer et al., 2007), e.g., 1.5M persons,

[1] https://lod-cloud.net/

810K places, 275K organizations: including 67K companies and 53K educational institutions (June 2019). LOD publishers can save a lot of time and effort by using RDF links from DBpedia for unambiguous 'real-world' objects, especially where dimensions are uncontentious, change only slowly, and do not affect interpretations from data, e.g., organizations, place, time. However, on the downside, for emerging data dimensions the linkability of LOD datasets need human intervention (Bryl et al., 2014) to help the 'web of data' playout. Despite the 1.5M people referenced in DBpedia, people connected to patent bibliographies and scientific publications remain an ambiguous data dimension where RDFs need disambiguation facilitated by human supervision.

In terms of patent related databases published by patent offices, there are only 2 such databases available: the one published by the European Patent Office (EPO)[2] and one by the Korean Patent office[3]. We focus on EP LOD database. The most recent LOD database, containing IPR data, is the Linked open EP data (henceforth: EP LOD), a raw data product provided by the EPO. The data is updated weekly. LOD is for now only marginally connected to other databases; in 2014 a test was done to explore its connection to DBPedia (the LOD version of Wikipedia), but has been dismantled since then (Kracker, 2017). Furthermore, links are potentially provided to other patent data, but only if it was published as linked open data by the relevant Patent offices, however, currently this is only the case for South Korea. Furthermore, there are some LOD databases provided by organizations and teams, which have transformed either the USPTO or EPO data into LOD (see e.g. Hassan et al, 2018).

Seeing the IP LodB map as a particular sub-cloud focusing on the patent and related innovation data, it is to note there are a few other sub-clouds with a particular foci in existence. The most known and documented (Belleau et al, 2008; Dumontier et al, 2011; Hu et al, 2017) subcloud is the Bio2RDF (https://bio2rdf.org/ and https://github.com/bio2rdf/bio2rdf-scripts/wiki). Bio2RDF is an open-source project that uses Semantic Web technologies to build and provide the largest network of Linked Data for the Life Sciences. Bio2RDF defines a set of simple conventions to create RDF(S) compatible Linked Data from a diverse set of heterogeneously formatted sources obtained from multiple data providers. It includes in Release 3, 35 datasets with approximately 11 billion triples.

Main axis short description

The main axis for alignment is the EP LOD- SN Sci Graph, however some additional alignments and their connected matchings are predicted (e.g. USPTO EP data).

EP LOD dataset contains the bibliographic data of EP applications and publications and their related applications / publications of other patent authorities. The data set covers EP applications[4] and EP publication data[5]. Bibliographic data on EP patents is quite comprehensive. Bibliographic data on non-EP patents is limited to basic patent identification information. In general the EP LOD[6] contains bibliographic data, e.g., patent number, date of filing (and granting), title, applicant data, inventor(s)' name, citation, classification and priority data, as well as patent family and cooperative patent classification (CPC) codes.

[2] Available at : https://www.epo.org/searching-for-patents/data/linked-open-data.html#tab-1

[3] Available at: http://lod.kipo.kr/

[4] EP applications, international applications (PCT applications) of EP applications, priorities of EP applications, non-EP applications which are in the same simple patent family (a.k.a. DOCDB family) as an EP application

[5] EP publications, publications which are cited by an EP publication, which are direct and indirect (indirect via Non Patent Literature) citations, non-EP publications whose application is in the same simple patent family (a.k.a. DOCDB family) as an EP application

[6] Available at: https://www.epo.org/searching-for-patents/data/linked-open-data.html#tab-1

The second axis, the Springer Nature (SN) SciGraph provides corroborating LOD focusing on scientific publication: articles, chapters, books, journals, people, grants, trials and patents (new in 2019). The Linked Open Data platform collates information from across the research landscape, for example funders, research projects, conferences, affiliations and publications, with more to follow and several yet realized other types of data.

The format of these databases allow that by following links, we discover more information, either within a certain database or between certain databases. These links can be pre-existing, and if the databases follow the standard an “same as” link is established between a record from one database to the record from another database. However, when establishing new links a more diligent approach will lead to truly identifying real-world objects in different databases.

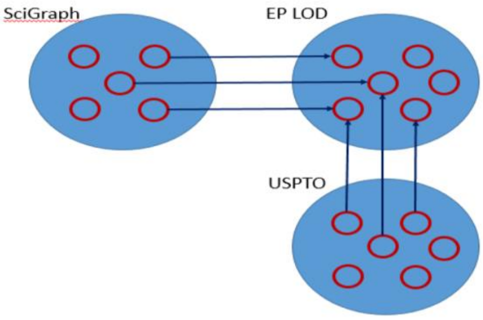


Figure 1. *Between databases linkages*

In terms of this main axis, we don’t see either EPO or SN as ‘more accurate’ than the other. We do however use a variety of auxiliary variables to help make these name and affiliation connections between the databases, e.g., geographic data. Nevertheless, for social science coding coherence between the source data and the ‘final’ coding is paramount, and the final coding may change as new data become available. Nonetheless, paramount to this additional data implementation is the discovery of relevant databases.

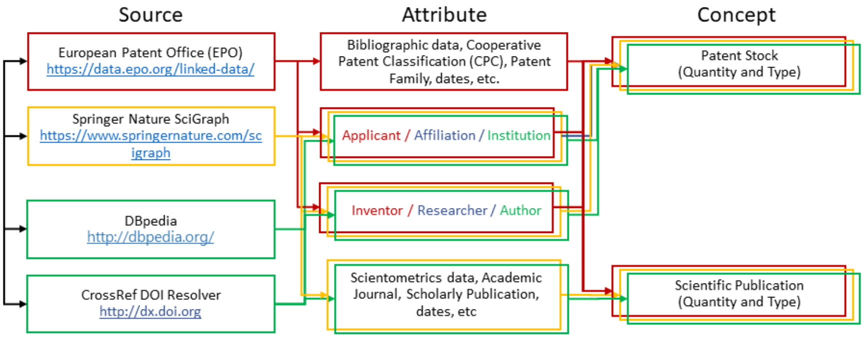

Source
Attribute
Concept
European Patent Office (EPO)
https://data.epo.org/linked-data/
Springer Nature SciGraph
https://www.springernature.com/scigraph
DBpedia
http://dbpedia.org/
CrossRef DOI Resolver
http://dx.doi.org
Bibliographic data, Cooperative Patent Classification (CPC), Patent Family, dates, etc.
Applicant / Affiliation / Institution
Inventor / Researcher / Author
Scientometrics data, Academic Journal, Scholarly Publication, dates, etc
Patent Stock (Quantity and Type)
Scientific Publication (Quantity and Type)

Figure 2. Disambiguation of actor names and institutional affiliations using data from two or more LOD sources. Source: Johnson et al. (2019)

Building the IP LOD Map

Discovery and Selection Workflow

The discovery and selection rely both on machine support (utilizing also both constructed tools), and a diligent scientific approach by the experts.

The process for database selection is a three step iterative process:

Step 1: Random fact gathering

Step 2: Technical screening

Step 3: Final expert selection

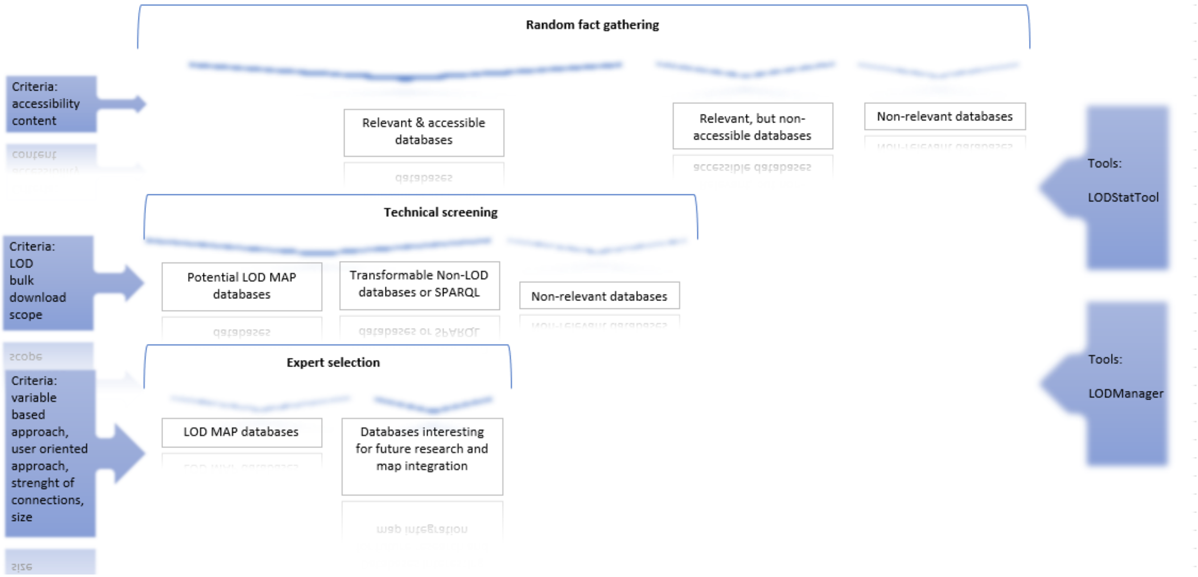

Random fact gathering
Criteria: accessibility content
Relevant & accessible databases
Relevant, but non-accessible databases
Non-relevant databases
Tools:
LODStatTool
Technical screening
Criteria: LOD bulk download scope
Potential LOD MAP databases
Transformable Non-LOD databases or SPARQL
Non-relevant databases
Expert selection
Criteria: variable based approach, user oriented approach, strenght of connections, size
LOD MAP databases
Databases interesting for future research and map integration
Tools:
LODManager

Figure 3. *Overview of the process for IP LOD map database selection (source: IP LODB project)*

Step 1 - Random fact gathering

Our literature review found several data aggregators or ‘information spaces’ (Dividino et al, 2014), from which discovery of LOD datasets related to IP is possible. The broadest of these aggregators is the LOD cloud. However, our experience using LOD Cloud shows that their hostings and aggregators suffer from lack of informative access information, high dataset dynamics, as well as low availability of resources (Umbrich et al, 2010; Dividino et al, 2014; Assaf et al, 2015).

When we checked links from LOD cloud (between June to October 2019) to original datasets we found that the majority of them are not available anymore (a problem also reported in the literature). However, some databases were later discovered on other servers. Hence, although we started by searching the LOD cloud, we proceeded by gathering additional information from several other data sources, e.g. Datahub, opendatahub, Manheim Linked Data Catalog, we also found significant number of additional active sources through recent scientific research articles that address the problems and benefits of LOD.

Originally, the datasets to be examined were divided between three team members and in this step we proceeded to categorize the LOD datasets (including some easily downloadable datasets that could be relevant) in three categories:

1. Relevant and accessible datasets (60 databases): Relevance was determined by the content assessment and all datasets that could potentially be aligned with the EP LOD, SN SciGraph or USPTO LOD dataset were put into this category. It was each team member's responsibility to also assess the accessibility at this point (following the links provided on the initial source site).
2. Relevant, but non-accessible datasets (15 databases): Our findings do not diverge from previous literature, warning about the problems of fragmented datasets availability and high level of flexibility in terms of their accessibility. All dataset data was re-examined trying to find if alternative hostings have been found separate to those provided by the initial datasource. If found, the dataset in question was re-qualified to "relevant and accessible datasets". However, we have been even in this initial step very restrictive in terms of naming a dataset "relevant" if it was not accessible, hence the low number of databases in this category.
3. Non-relevant databases (130 databases): The rest of the databases that were at least somewhat interesting are put in this last category, however they are omitted from further examination.

Step 2 - Technical screening

Technical screening builds upon the relevant and accessible databases and also included the step where we tested the transformation efforts needed to turn a non-LOD database into a LOD (see more under transformable non-LOD databases). It includes the re-assessment of databases, partial use of the LODStatTool (when needed) and re-checking of accessibility.

In this step the databases were re-examined and classified in the following way:

1. Potential LOD MAP databases: A joined effort by 4 team members. The step involved re-investigating the databases also to classify them into several types of data collection: i. Patent databases [Pat]; ii. Publication databases [Pub]; iii. Technology databases [Tech], which includes downstream data, i.e., databases on products; iv. Grants databases [Grant], which includes upstream data, i.e. data on competitive public grant funding; v. Databases containing information on entities [Entities], including predominantly data on organizations, but with some data connected to individuals, vi. Databases containing geographical data [Geo], vii. Wiki databases [Wiki], containing data for verification, viii. Circular Economy related databases, which are useful mainly later on for exploring a more focus-oriented stream (CE), and ix. Miscellaneous databases [Misc], which are useful in a widened scope. 40 databases remain in this category.

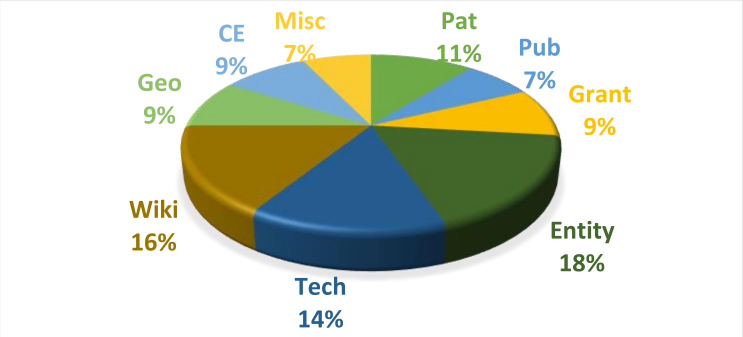


Figure 4. *Distribution of databases by collection type (source: IP LODB project)*

There are diverse benefits to connecting patent and publication data, as well as connecting it to other types of databases. Please note that the connection between patent and publication data is amply explained in 3.1, where benefits of connecting these two types of databases is also explained. The benefit of connecting to geo databases is to either enhance missing data or provide validation of data for individual records (e.g. enhancing locational data with exact locations, either in terms of addresses or longitudinal and latitudinal data) allowing for easier, more precise disambiguation, or to be able to provide additional input for analysis focusing on geography of innovation (e.g. finding geolocational data for bursts of new technologies which are promoted by the governments, e.g. Circular Economy[7]; or for new emerging technologies). Furthermore, connecting the data both upstream (to research) and downstream data (to products) is important. There is a lack of product base databases, however these are reachable especially for pharma products due to stringent procedures for bringing this to market. Please note, we did not include a recent effort of one of the ARP EPO projects (from the first year of the ARP grant, the Insights from product-patent correspondence), since no data is available up to this point (see http://www.iproduct.io/data). We otherwise find that dataset as potentially a very valuable add-on to the map, especially if the dataset is either originally available in LOD or would be transferred into LOD format. Quality data on entities is also rare; albeit there is no lack of partial or less-quality datasets. We found GRID[8] to be the most comprehensive and reliable entities database. Furthermore, the DBPedia is the most linked data, with Wikidata providing a more comprehensive opportunity for data that can provide context data, additional explanations of terminology etc. However, since the EPO LOD has already tested this link to DBPedia (Kracker, 2017), we have not engaged in a more comprehensive linking, since we hope that sometime in the future more information will be available to us on this previous effort. Lastly, the grants data has become a very comprehensive data source. Grant data can enlighten not only additional cooperation efforts, but also help track the effect of public funding. Currently, we already connect records (including providing links) to Springer Grant data, but plan to invest further efforts into connecting also other databases containing data on public grants (e.g. Open Aire is of particular interest).

2. Transformable non-LOD databases or SPARQL: Some datasets were found to be relevant, but were not in one of the LOD formats. However, there was a bulk download available, hence we proceeded to test if these datasets could be transformed to LOD by the fourth member of the team or if their information could be added at a later stage (presumably post-project). The test (briefly reported upon in the preliminary report) discovered this could be done without significant efforts as there are several options available for how to proceed. We also used this knowledge to later design the IPLOD bulk download.
3. Non-relevant databases: Databases deemed not (contextually) relevant were moved to this type.

Step 3: Expert selection

[7] Note due to our special focus interest in Circular Economy, we also include some datasets which provide data on green technologies and their dispersion.

[8] GRID (available at: https://www.grid.ac/) includes information on almost 100,000 organizations, out of which about 30% are companies, 20% are higher education institutions (HEI), with about 10% nonprofit and 10% hospitals. The database includes several variables, such as address, type, the URl of the organization etc. (GRID, 2020).

The aim of this step is to get a selection of LOD databases, which would be explored more in-depth. In order to do so, we have prioritized the databases and the databases containing the Priority 3 or 4 will be the ones we will continue working with. Priority 4 represents the main axis: EP LOD (and Springer) and the SN SciGraph. The main criteria to determine priority 3 was, whether the data contains patent information (patent numbers or scientific publications) or other relevant data which can also be directly linked to patents or SN articles and the information available from one or the other Priority 4 database. An iterative approach, were initial prioritizations were re-examined various times by different team members (regardless of the person who assigned the priority initially) brought us to a selection of 12 databases as our top priority. These priority databases have either already been interlinked inside the IP LodB database (e.g. GRID, Springer Grants), are ready for the link to be enforced after resolving issues (e.g. GeoNames, where link will be only enforced when shifting to other types of entities, i.e. organizations), or the links have been explored in-detail in regard to their connectivity with EPO (e.g. KIPO LOD or UNIPROT).

Determination of Links Workflow

Having selected the databases that will be included in the IP LODB map, we then proceed by examining the (existing and) potential links – where we limit this to those deemed most important. The determination of links proceeds in various steps:

Step 0 – Testing

We first examined the LOD Cloud information for Priority 3 and Priority 4 databases that are included in the cloud. Next, we examine the literature and other available information in regard to these databases. Furthermore, we examine the ontologies of these selected databases. This gave us an matrices of potential alignment potential between pairs of databases. Furthermore, the testing led to understanding the necessary next steps and tasks: a) understanding the potential and existing linkages discovery with identification of same real-life objects in pairs of databases and determination of potential (and existing) linkages; b) vocabulary alignment and ontologies identification; c) identification of auxiliary variables together with index variables to align on and construction of links.

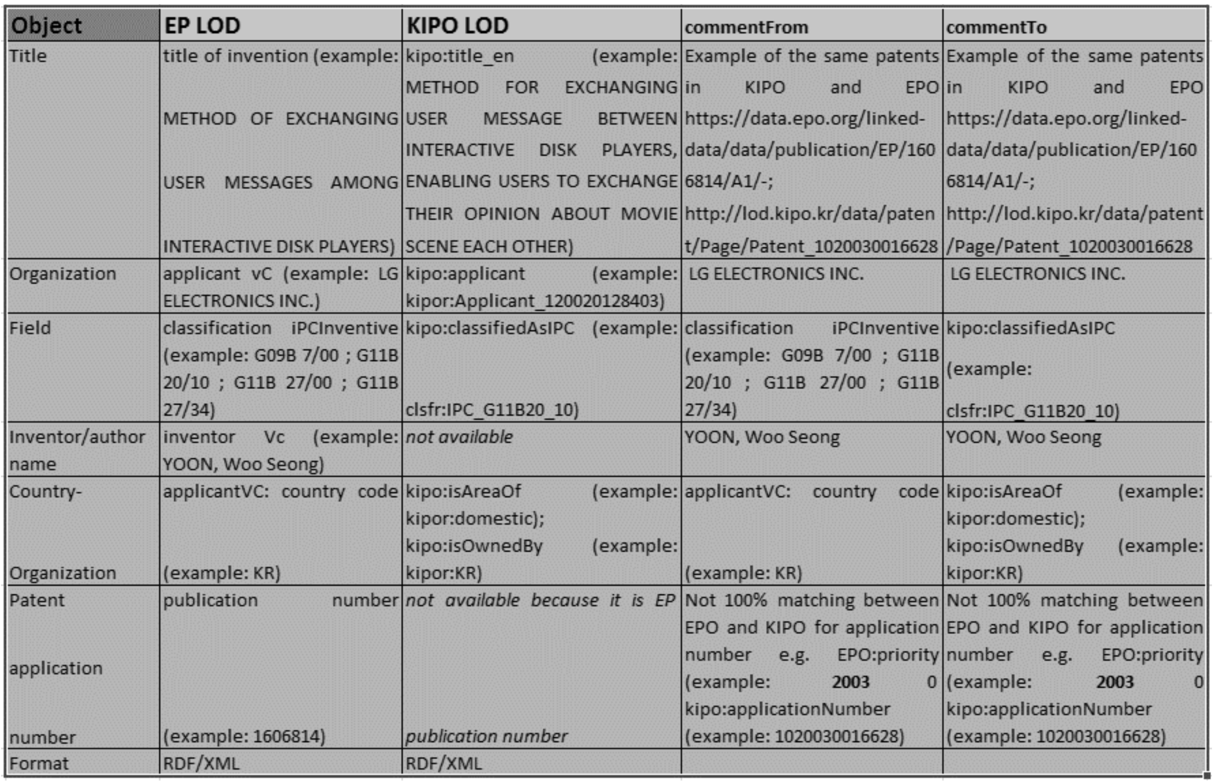
Step 1 – Initial potential and existing linkages discovery

A matrix guided the construction of a complete table with alignment pairs which points out to concrete *index variables* used to align on. This was later used as an input to the visualization of the IP LOD cloud for the selected databases.

Step 2 – Ontologies identification and vocabulary alignments

One of the often reported problems of the LOD datasets are the diverse ontologies and the vocabulary mismatches. We re-construct the ontologies to built a cross-table with variable names for identified potential pairings for relevant databases.

Table 1. *Cross-table (EP-KIPO example)*



Object	EP LOD	KIPO LOD	commentFrom	commentTo
Title	title of invention (example: METHOD OF EXCHANGING USER MESSAGES AMONG INTERACTIVE DISK PLAYERS)	kipo:title_en (example: METHOD FOR EXCHANGING USER MESSAGE BETWEEN INTERACTIVE DISK PLAYERS, ENABLING USERS TO EXCHANGE THEIR OPINION ABOUT MOVIE SCENE EACH OTHER)	Example of the same patents in KIPO and EPO https://data.epo.org/linked-data/data/publication/EP/1606814/A1/-; http://lod.kipo.kr/data/patent/Page/Patent_1020030016628	Example of the same patents in KIPO and EPO https://data.epo.org/linked-data/data/publication/EP/1606814/A1/-; http://lod.kipo.kr/data/patent/Page/Patent_1020030016628
Organization	applicant vC (example: LG ELECTRONICS INC.)	kipo:applicant (example: kipor:Applicant_120020128403)	LG ELECTRONICS INC.	LG ELECTRONICS INC.
Field	classification iPCInventive (example: G09B 7/00 ; G11B 20/10 ; G11B 27/00 ; G11B 27/34)	kipo:classifiedAsIPC (example: clsfr:IPC_G11B20_10)	classification iPCInventive (example: G09B 7/00 ; G11B 20/10 ; G11B 27/00 ; G11B 27/34)	kipo:classifiedAsIPC (example: clsfr:IPC_G11B20_10)
Inventor/author name	inventor Vc (example: YOON, Woo Seong)	*not available*	YOON, Woo Seong	YOON, Woo Seong
Country-Organization	applicantVC: country code (example: KR)	kipo:isAreaOf (example: kipor:domestic); kipo:isOwnedBy (example: kipor:KR)	applicantVC: country code (example: KR)	kipo:isAreaOf (example: kipor:domestic); kipo:isOwnedBy (example: kipor:KR)
Patent application number	publication number (example: 1606814)	*not available because it is EP publication number*	Not 100% matching between EPO and KIPO for application number e.g. EPO:priority (example: **2003** 0 kipo:applicationNumber (example: 1020030016628)	Not 100% matching between EPO and KIPO for application number e.g. EPO:priority (example: **2003** 0 kipo:applicationNumber (example: 1020030016628)
Format	RDF/XML	RDF/XML		

Step 3 – Construction of links for IPLOD map

With construction of links for the IPLOD map we wanted to present pairs of databases which are somehow related to each other and basic information in databases that presents connectivity between them. To do that we first defined databases of interest, based on content description for each of the databases, after that we defined means of connectivity for pairs of databases. Means of connectivity was defined as a set of properties in each of pairs of databases that could represent the same object in both databases. For instance databases EP LOD and KIPO which are both patent databases could contain the same patent numbers, authors, affiliations and descriptions. Database Uniprot is a technology database containing information about proteins and could be related with EP LOD database through patent numbers. For each pair of databases we defined actual or possible means of connectivity described as a sentence.

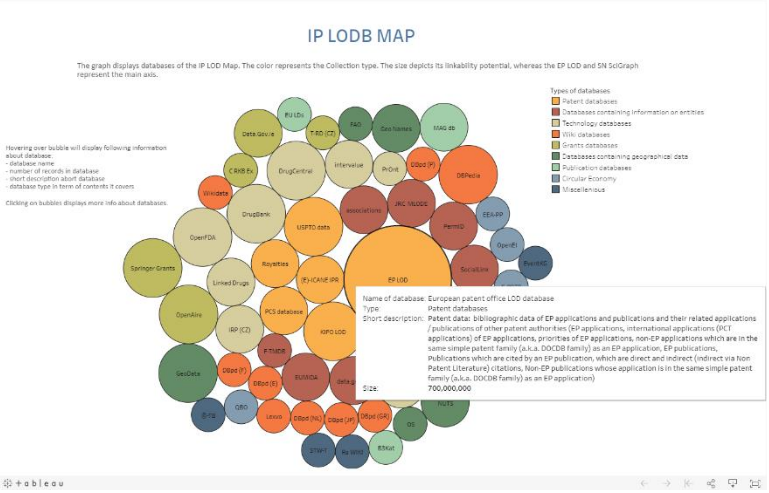
Visualisation of the Map

Visualization of LOD is important both in terms of dataset discovery, for supporting the dataset selection process (Papadaki et al, 2018) and for allowing non-domain and non-technical audiences to obtain a good understanding of the structure and contents published in LOD, and be able to compose queries, identify links between resources and intuitively discover new pieces of information (Dadzie and Rowe, 2011).

We explored several options for visualization of the IP LOD map(s). The most known visualization (diagrams) of the LOD databases, the LOD cloud, have been done by LOD Cloud Draw[9]. There were also other initiatives upgrading this to 3D visualizations (e.g. Papadaki et al, 2018) or LDVP, and seeking for more interactive approaches. However, based on comparison analysis, our current knowledge and experience with R tools and Tableau we decided to use a combination of Tableau and then R library for visualization and the R Shiny framework for sharing interactive visualizations via web page.

Tableau products query relational databases, online analytical processing cubes, cloud databases, and spreadsheets to generate graph-type data visualizations. The products can also extract, store, and retrieve data from an in-memory data engine. Rshiny is a server environment which runs R scripts and enables inclusion of various types of controls such as text inputs, dropdown menus, hover and highlight over graphical objects, clicks on graphical objects with full integration of R environment.

The process is multilevel. For the *basic IP LODB map* the source data is prepared, containing all 40 databases (from step 2) with their selected attributes (description, size, etc.). The map is embedded in the webpage iplod.io and all the additional information is available in the two maps provided there.



IP LODB MAP

The graph displays databases of the IP LOD Map. The color represents the Collection type. The size depicts its linkability potential, whereas the EP LOD and SN SciGraph represent the main axis.

Hovering over bubble will display following information about database:
- database name
- number of records in database
- short description abort database
- database type in term of contents it covers

Clicking on bubbles displays more info about databases.

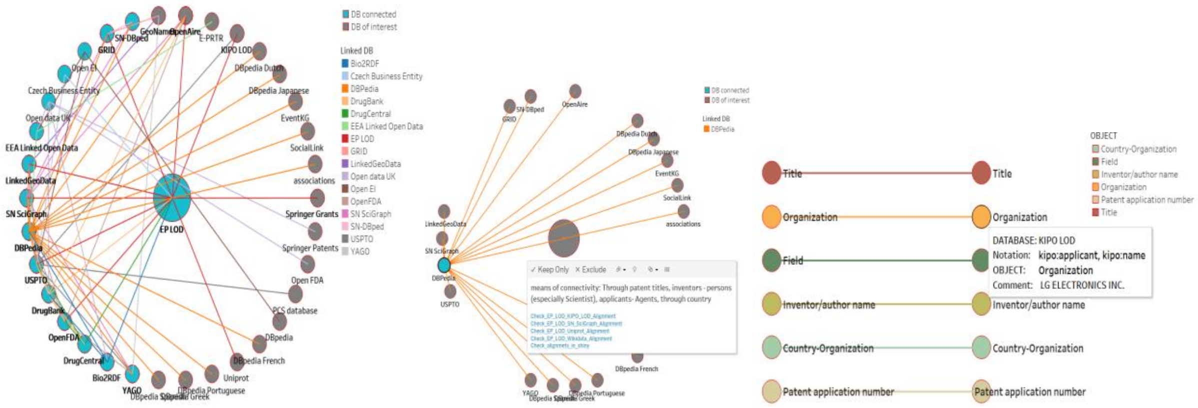
Name of database: European patent office LOD database
Type: Patent databases
Short description: Patent data: bibliographic data of EP applications and publications and their related applications / publications of other patent authorities (EP applications, international applications (PCT applications) of EP applications, priorities of EP applications, non-EP applications which are in the same simple patent family (a.k.a. DOCDB family) as an EP application, EP publications, Publications which are cited by an EP publication, which are direct and indirect (indirect via Non Patent Literature) citations, Non-EP publications whose application is in the same simple patent family (a.k.a. DOCDB family) as an EP application)
Size: 700,000,000

Figure 5. *IP LODB Map – basic MAP version*

Visualizations of the links with Tableau

In Tableau in order to express basic links between different databases we used a star network layout with EP LOD as the central database, and remaining databases distributed around it.

[9] https://github.com/lod-cloud/lod-cloud-draw. The solution is written in Rust, can be compiled with Cargo and creates a LOD cloud diagrams as SVG.



means of connectivity: Through patent titles, inventors - persons (especially Scientist), applicants- Agents, through country

Check_EP_LOD_KIPO_LOD_Alignment
Check_EP_LOD_SN_SciGraph_Alignment
Check_EP_LOD_Uniprot_Alignment
Check_EP_LOD_Wikidata_Alignment
Check_alignments_in_shiny

DATABASE: KIPO LOD
Notation: kipo:applicant, kipo:name
OBJECT: Organization
Comment: LG ELECTRONICS INC.

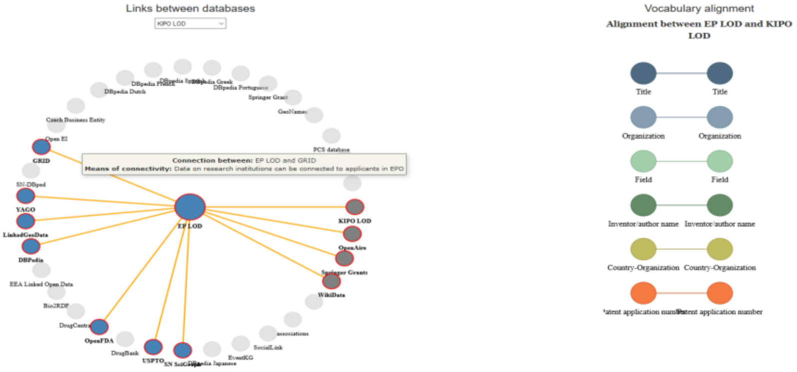
Figure 6. *Links and example of the vocabulary alignment*

Left you can see the links between databases in star layout. In the center you can observe links between DBPedia and the hover over link displaying means of connectivity and right is an example of vocabulary alignment between EP LOD and KIPO LOD databases.

By clicking on one of databases only links to the clicked database can be shown, by hovering over link between two databases information about possible parameters for connectivity are displayed as well as links to vocabulary alignments for specific databases. Vocabulary alignment for a pair of two specific databases can be accessed via the click on menu that appears after hovering over the link between two databases. In vocabulary alignment instances of objects that can be used as a mean of connectivity between two databases are displayed.

Visualization of the links with R Shiny

We also made a visualization of links between databases and vocabulary alignment in R Shiny environment with the use of VisNetwork library for R. VisNetwork library for R is a R implementation of JavaScript library VisNetwork which enables interactive development of network graph visualizations in R environment. The main difference from visualization in Tableau is that instead of switching between different pages for vocabularies for pairs of specific databases we did that directly within one page with the same functionality as visualization in Tableau.



Connection between: EP LOD and GRID
Means of connectivity: Data on research institutions can be connected to applicants in EPO

Figure 7. *Example of visualization with RShiny.*

Dataset diagnostics

We implement a ping to home links and SPARQL endpoints, inspired by our practical observations, expert recommendations and works such as Umbrich et al (2010). Some datasets, such as DBpedia, have been available from the very beginning of the LOD movement and regularly undergo changes on both the instance level (triples) and the schema level (classes). New resources are added and old resources are removed; new links are set to other datasets, and old links are removed as the target has vanished. We can safely say that the datasets in the LOD cloud are dynamic (Umbrich et al, 2010).

About two thirds of checked databases offer access to LOD data through SPARQL endpoints. By pinging SPARQL endpoints we wanted to check whether endpoints are alive, number of OWL classes contained in data and number of triples in specific databases. Apart from that we checked for availability of home page links for all databases by checking http response codes for each database.

SPARQL endpoints could be accessed via HTTP or HTTPS protocol and are thus easily available through a variety of libraries and tools available in different programming languages. For basic ping functionality test, a simple batch script is sufficient, but if we want to analyze results of sparql queries, tools and libraries that implement SPARQL protocol should be used. Such libraries wrap SPARQL requests, send them to the server, and provide means to easily parse responses from SPARQL server. In our case SPARQL ping request functionality is implemented with Apache Jena Java library which is a free and open source library for building Semantic web and Linked data applications https://jena.apache.org/.

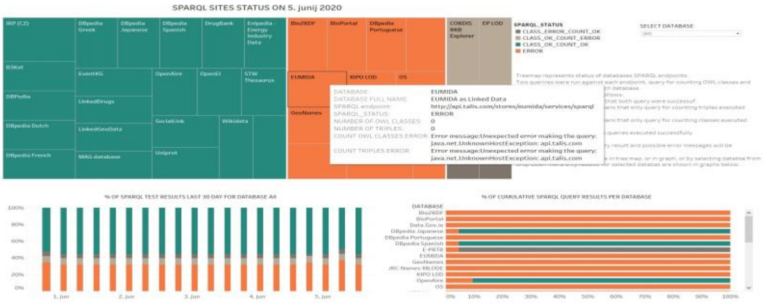
For the basic ping test we first try to connect to each database home link by calling basic http request to the home database's url. We record responses in terms of http response codes, for instance 200 OK means that link is alive, 404 Not Found means that the url address on which we tried to connect is not found on given address. In case there is no response we record a Java exception message. For SPARQL endpoint we first send a query which lists all rdf classes followed by a query which counts all triples in the database. By counting rdf classes we wanted to check whether structure of the database is changing over time, by counting triples we wanted to check whether the database is being updated with new data.

Testing SPARQL endpoints

Results for http ping and SPARQL test are presented as a Tableau dashboard where we marked whether the home link for the database site was alive on the last test run. For sites where sparql endpoint is available, we marked whether both of queries executed successfully, for queries that were not executed successfully we presented reasons in terms of Java exception messages. Results of ping tests and tests of SPARQL endpoints are presented as treemaps with two additional graphs with statistics for each test run and cumulative statistics for each database over all test runs.

For representing results of SPARQL endpoint tests we used treemap charts which represent data in hierarchical groups of nested rectangles where size of rectangles can represent one of measures in result data. For grouping variable we defined a variable SPARQL_STATUS with following four options as result of two sparql query types:

- **CLASS_OK_COUNT_OK**: means that both queries were successful.
- **CLASS_ERROR_COUNT_OK**: means that only query for counting triples were executed successfully.
- **CLASS_OK_COUNT_ERROR**: means that only query for counting classes were executed successfully.
- **ERROR**: means that none of the queries executed successfully.



DATABASE: EUMIDA
DATABASE FULL NAME: EUMIDA as Linked Data
SPARQL endpoint: http://api.talis.com/stores/eumida/services/sparql
SPARQL_STATUS: ERROR
NUMBER OF OWL CLASSES: 0
NUMBER OF TRIPLES: 0
COUNT OWL CLASSES ERROR: Error message:Unexpected error making the query: java.net.UnknownHostException: api.talis.com
COUNT TRIPLES ERROR: Error message:Unexpected error making the query: java.net.UnknownHostException: api.talis.com

Treemap represents status of databases SPARQL endpoints.

***Figure 8:** Example of dashboard presenting results of SPARQL endpoint test.*

In the case of SPARQL endpoint tests treemap shows only data for the most recent test run, history results are represented in two other charts.

Some measured values had value zero or number of classes or triples for some databases varied significantly or could not be acquired. For instance UniProt database contains over 66 billion triples, database DrugBank contains only half a million triples while with EP LOD database we were unable to count the number of triples, in such cases values were set to 0. Regarding measuring response times in SPARQL requests, particularly for counting triples, we have to note that some databases contain prestored values for triples count, and such queries require less than one second to execute, while some databases perform actual triples count which strongly depends on database size. In the worst case triples count does not execute or takes up to 10 minutes to execute, while some triples count takes milliseconds to execute. Such variations in values would not represent any informative value.

SPARQL endpoints testing results

We have started a stable testing from June 1st. Prior attempts were more sporadic in nature, hence we do not include any results from the period before.

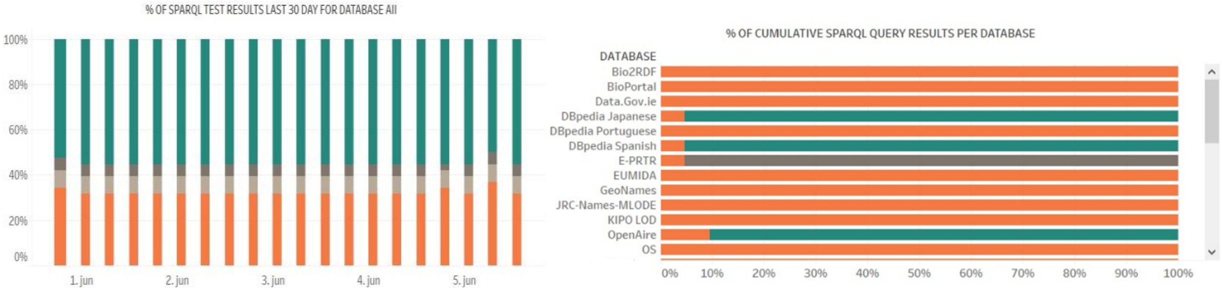
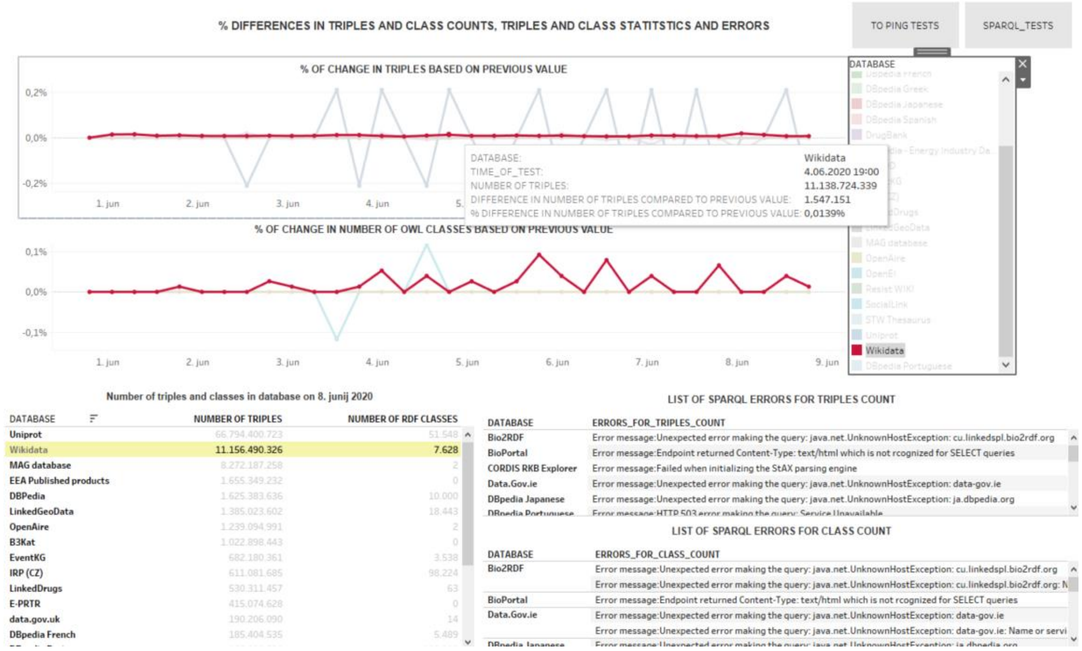


***Figure 9.** Testing results*

Left are results for specific test run expressed in percentage of SPARQL_STATUS result from all databases under test; right are cumulative results for each database for all test runs with percentage for SPARQ_STATUS result.

As the number of triples or OWL classes in some databases is really huge compared to others where some databases contain only a few thousand records, daily changes in the number of triples of a couple hundred thousands over a billion already existing in the database would hardly be noticed. Thus we represent change of number of classes and triples by representing it as percentage change for each database compared to the value in the previous test run. In such a way dynamics for data update and data structure change for each database could be observed over time.

Our data shows (unsurprisingly) that Wikidata is the most active LOD database, which contains relevant records which can provide additional value to innovation data. Changes are detected on both the instance level (triples) and the schema level (classes).



DATABASE:	Wikidata
TIME_OF_TEST:	4.06.2020 19:00
NUMBER OF TRIPLES:	11.138.724.339
DIFFERENCE IN NUMBER OF TRIPLES COMPARED TO PREVIOUS VALUE:	1.547.151
% DIFFERENCE IN NUMBER OF TRIPLES COMPARED TO PREVIOUS VALUE:	0,0139%

Number of triples and classes in database on 8. junij 2020

DATABASE	NUMBER OF TRIPLES	NUMBER OF RDF CLASSES
Uniprot	66.794.400.723	51.548
Wikidata	11.156.490.326	7.628
MAG database	8.272.187.258	2
EEA Published products	1.655.349.232	0
DBPedia	1.625.383.636	10.000
LinkedGeoData	1.385.023.602	18.443
OpenAire	1.239.094.991	2
B3Kat	1.022.898.443	0
EventKG	682.180.361	3.538
IRP (CZ)	611.081.685	98.224
LinkedDrugs	530.311.457	63
E-PRTR	415.074.628	0
data.gov.uk	190.206.090	14
DBpedia French	185.404.535	5.489

LIST OF SPARQL ERRORS FOR TRIPLES COUNT

DATABASE	ERRORS_FOR_TRIPLES_COUNT
Bio2RDF	Error message:Unexpected error making the query: java.net.UnknownHostException: cu.linkedspl.bio2rdf.org
BioPortal	Error message:Endpoint returned Content-Type: text/html which is not rcognized for SELECT queries
CORDIS RKB Explorer	Error message:Failed when initializing the StAX parsing engine
Data.Gov.ie	Error message:Unexpected error making the query: java.net.UnknownHostException: data-gov.ie
DBpedia Japanese	Error message:Unexpected error making the query: java.net.UnknownHostException: ja.dbpedia.org
DBpedia Portuguese	Error message:HTTP 503 error making the query: Service Unavailable

LIST OF SPARQL ERRORS FOR CLASS COUNT

DATABASE	ERRORS_FOR_CLASS_COUNT
Bio2RDF	Error message:Unexpected error making the query: java.net.UnknownHostException: cu.linkedspl.bio2rdf.org
	Error message:Unexpected error making the query: java.net.UnknownHostException: cu.linkedspl.bio2rdf.org: N
BioPortal	Error message:Endpoint returned Content-Type: text/html which is not rcognized for SELECT queries
Data.Gov.ie	Error message:Unexpected error making the query: java.net.UnknownHostException: data-gov.ie
	Error message:Unexpected error making the query: java.net.UnknownHostException: data-gov.ie: Name or servi
DBpedia Japanese	Error message:Unexpected error making the query: java.net.UnknownHostException: ja.dbpedia.org

Figure 10. *Wikidata test for triples and classes*

Dashboard with percentage change in number of triples and OWL classes, recent values for triples and number of classes, and errors for specific databases with highlighted value for Wikidata. Interactive version is available on iplod.io webpage.

Not only can we observe differences in the dynamics in terms of percentage changes between databases, but we can also observe different time dynamics. We can observe that on Wikidata new triples are added on a daily basis, while for Uniprot a new content was added at the end of June. In database E-PRTR also some negative trends can be observed which shows that data is also being deleted.

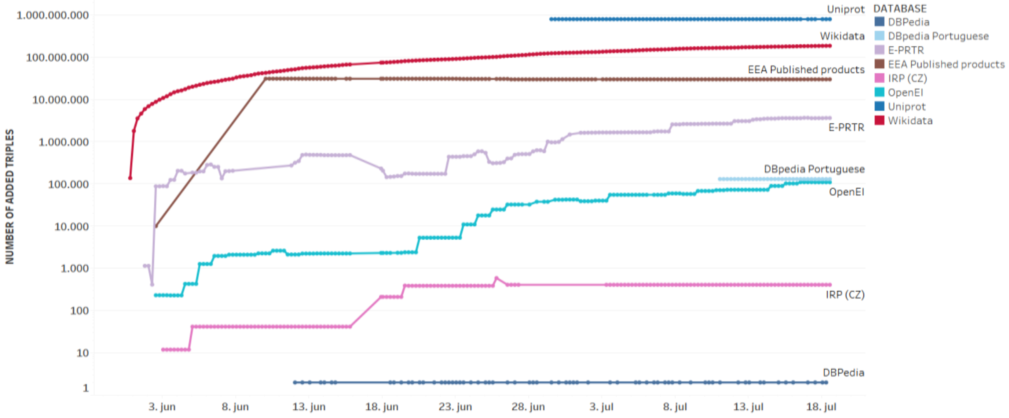


Figure 11: *Databases new triples (instances level)*

Cumulative number of added new classes over time for most active databases with changes in number of triples at least 1000 between consecutive test runs. In order to show activity on all databases we are using logarithmic scale since changes in some databases are very small compared to other databases. Lines on *graph start when new triples are added for a given database. Please note that for EP LOD database we could not count number of triples since database time out before the query could execute and Springer does not provide a SPARQL endpoint.*

We can also observe cumulative number of triples for all databases that are adding new content. Here we can observe that In some databases content is being removed for instance in E-PRTR, OpenEI, EEA Published products and IRP(CZ) where number of removed triples is being less than 100 triples. The DBPedia lines (bellow in green) show an interesting trend.

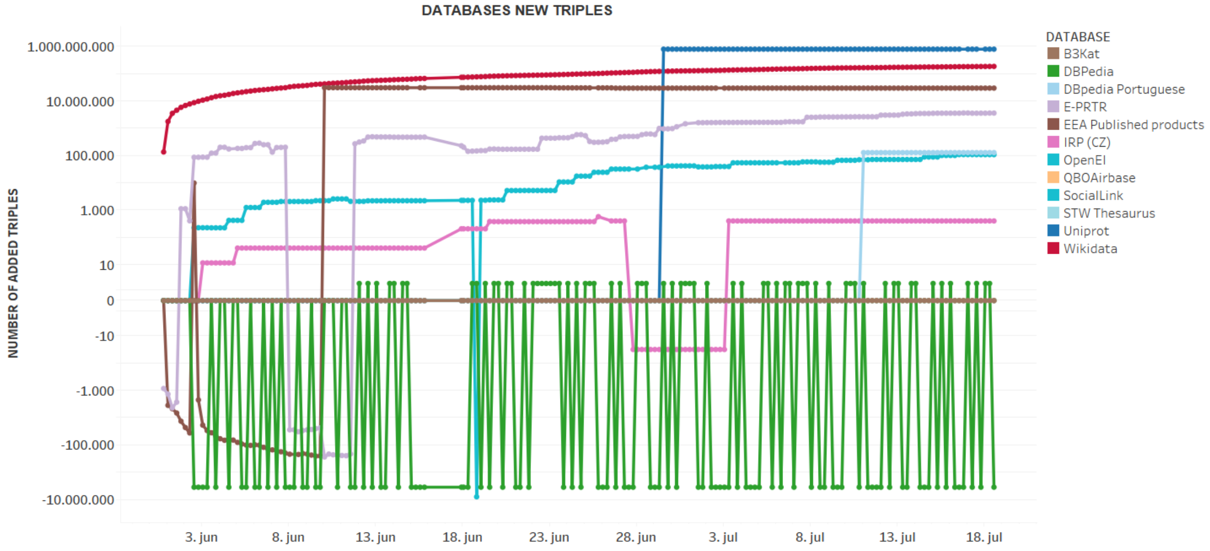


***Figure 12**. Cumulative number of triples (instances level)*

With DBPedia database that same amount of triples is being added and removed between consecutive test runs. In order to inspect the activity a detailed insight in content would be required.

As said we can also observe how new classes are added to each database. Graph starts when new class was added to existing database structure. New classes means that new properties are appearing in database which suggests that database is in active use.

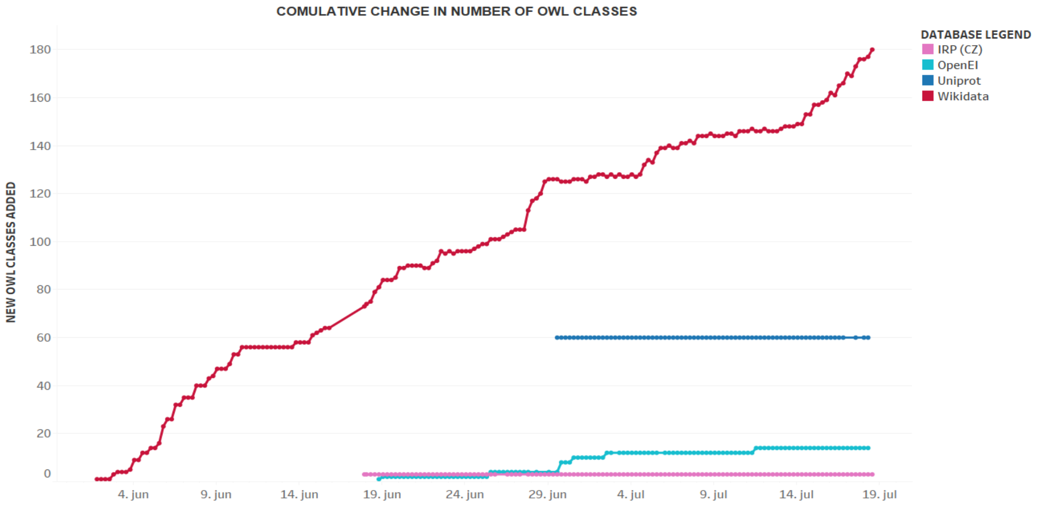


Figure 13. Shema level (triples)

For Wikidata we can observe that new classes and added on daily bases since content in Wikidata is being changed on daily basis, while Uniprot database receives new data on monthly basis where with new data for proteins also new classes appear.

SPARQL endpoint test results overview

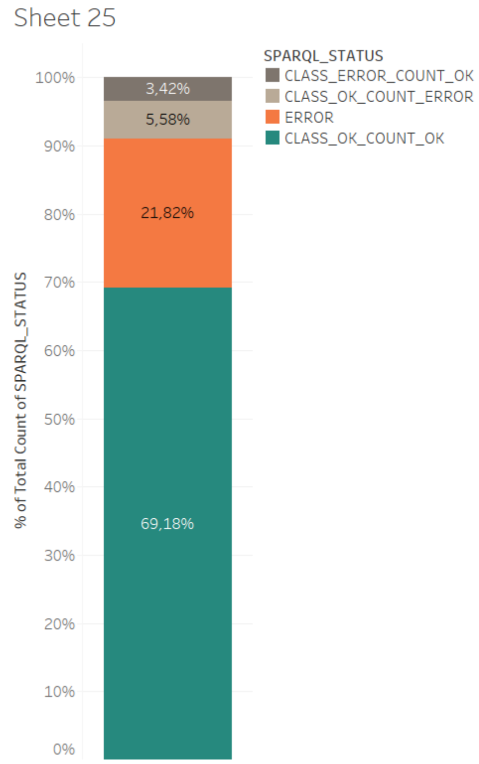


Figure 14. *SPARQL endpoint test results statistics*

Overall 69,16 % of SPARQL endpoint test runs executed successfully, in 5.58% only count of triples was not executed successfully, in 3.42 % only count of OWL classes was not executed successfully and in 21.82% none of sparql requests was executed successfully (period June 2020 - Jan 2021). The main reason that count of classes was not executed successfully was that endpoint or proxy server in front of endpoint signaled timeout by sending HTTP 503 status code Service Unavailable, such databases are EP LOD and E-PTR. In case of Uniprot database query for count of triples returned result and with redirect command for new query which did not pass through Apache Jena client syntax check, even the fact that query worked properly on web page, but after we slightly modified, modified version of query worked properly. We noticed that some databases have a prestored number of triples and results are returned instantly, while others perform actual count and in case of large number of triples such operation may take significant amount of time in case of larger databases. Prestored values for number of triples are available in databases of vendor Virtuoso https://virtuoso.openlinksw.com/ which is used by DBPedia databases and Uniprot database, and Blazegraph https://blazegraph.com/ used by Wikidata. For many endpoints we were unable to determine the vendor since home links from which SPARQL queries are sent do not reveal database vendor information.

The main reason that none of the queries worked or were unsuccessful was either the fact that SPARQL endpoint listed on site was down or results returned from endpoint were not in correct format. For the case when endpoint was down the following Java exception messages were shown:

- Connection to url failed.
- Connection by url was refused.
- UnknownHostException in case that URL name was not resolved by DNS.

***Table 2**. Examples of SPARQL errors related to connectivity.*

Example of connectivity error	Explanation
Error message:Unexpected error making the query: java.net.UnknownHostException: cu.linkedspl.bio2rdf.org	Here URL **cu.linkedspl.bio2rdf.org** could not be resolved by the DNS server, hence connection to it could not be made.
Error message:Unexpected error making the query: org.apache.http.conn.HttpHostConnectException: Connect to mlode.nlp2rdf.org:80 [mlode.nlp2rdf.org/139.18.2.138] failed: Connection timed out (Connection timed out)	Here URL to mlode.nlp2rdf.org was resolved, but connection to the IP address 139.18.2.138 and port could not be established within time frame before client times out.
Error message:Unexpected error making the query: org.apache.http.conn.HttpHostConnectException: Connect to lod.cs.aau.dk:8891 [lod.cs.aau.dk/192.38.56.34] failed: Connection refused (Connection refused)	Here the server on IP address 192.38.56.34 refuses connection to the 8891.

The second most important reason that the query was not executed successfully was due to error in format or content type, also known as MIME type, in which the endpoint returned results to the Apache Jena client or data in response could not be parsed by Apache Jena client. This was usually the case with endpoints that were meant to be used directly through home link websites or were set by default to return results in HTML format. Such format are meant to be presented directly on the web pages for testing or development purposes while SPARQL clients naturally support formats like n-triples, RDF/XML or JSON. For some endpoints we managed to modify the query by adding appropriate parameter to be either SPARQL or JSON in the endpoint's URL, for instance http://wiki.rkbexplorer.com/sparql/?**format=sparql&**. We have to note that adjusting format parameters was not available for all SPARQL endpoints and is not always available under format name and neither is available on all endpoints. Option for adjusting format parameters depends on SPARQL endpoint implementation and properties settings for related endpoints, usually if it is available, it is on html sites where SPARQL queries could be tested where users can generate HTTP GET by clicking on endpoint parameters, like on DBPEDIA http://dbpedia.org/sparql. Some endpoints returned values in RFD/XML or JSON format that could not be parsed by Apache Jena SPARQL client.

***Table 3.** Examples of SPARQL errors related to format or MIME type and errors in endpoint's response.*

Example of MIME type error	Explanation
Error message:Endpoint returned Content-Type: text/html which is not recognized for SELECT queries. (BioPortal)	Servers returns result as HTML page which is not recognized or accepted by Apache Jena SPARQL client.
Error message:Failed when initializing the StAX parsing engine	StAX parsing engine exception is returned by Apache Jena SPARQL client due to errors in returned data from endpoint.
Error message:One or more of the required keys [head, results] was not found	Key head and results were missing in response from SPARQL endpoint and results couldn't be read by client.

The third reason that SPARQL query could not be executed correctly was the fact that the query was timed out or was not executed correctly due to some internal error on SPARQL endpoint. Errors for such queries were accompanied with HTTP error code 503 Service Unavailable or 502 Proxy Error or code 500 Internal server error. When HTTP code 503 is returned this usually means that requested service is not operational but in order to inspect reasons properly a response from SPARQL endpoint administrator would be required. It is quite often that error codes on http servers could be modified by administrators or developers, hence the same error codes could be returned for different reasons. We noticed that in case of query timeout for database EP LOD error code 503 is returned after 90 seconds when triples count query was sent, even the fact that query for counting classes executed without problems.

SPARQL endpoint test final remarks

Usage of sparql endpoints through the web pages is a natural choice for many users, particularly when someone wants to make a quick search for some data in a database or tests query for later use on dedicated software. Our goal was to test whether SPARQL endpoints are alive and if they return data for specific queries.

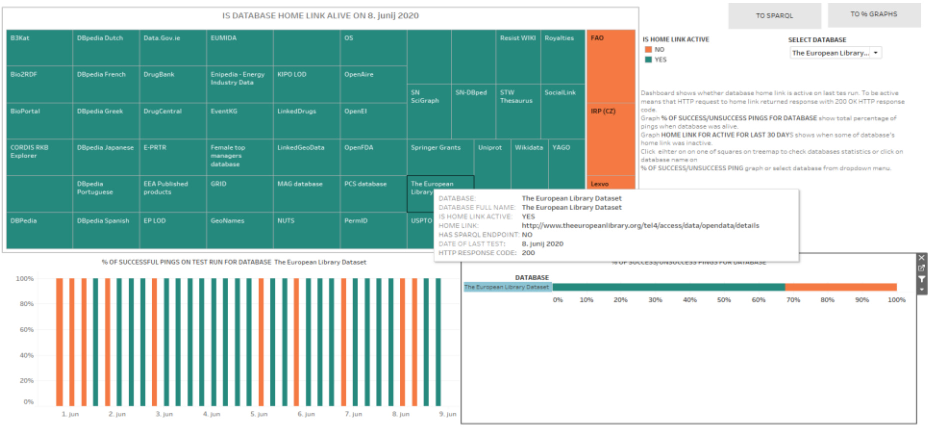
Based on results we can say that many endpoints work sufficiently well for basic graph searches with limited amount of triples pattern in WHERE condition and exact filters. For complex triples pattern in WHERE condition, and in case of aggregate SPARQL queries query time can rise significantly. Databases which timeout on triples count would definitely not be able to execute query on time, such example is EPO LOD. Such complex queries should be performed by the end user locally which takes resources in terms of computing power, time on knowledge about administering SPARQL endpoints and triples stores.

For that many open source or to some extent free SPARQL databases exist https://en.wikipedia.org/wiki/List_of_SPARQL_implementations, almost all endpoints under test provide dumped contents of databases to be downloaded and stored on local SPARQL databases. With that no extra load is put on public databases for complex queries, and there is no need to transfer large data volumes over the internet in case of querying multiple triple stores simultaneously.

Ping to databases home link result

For presenting ping to databases home link we used the same logic as we did with the ping to SPARQL endpoints except that for presenting the result we defined variable IS_HOME_LINK_ACTIVE. In case that http get connection attempt to databases home link site returns 200 OK HTTP response code we set variable's value to YES, in case that other HTTP response codes are returned, particularly codes 4xx or 5xx, or Java exception message is thrown, value is set to NO.

Generally only three home links are constantly down, others are alive with only exceptions when sites are temporarily down due to maintenance. Result for most recent status is presented in a treemap chart, results for each test run are presented with stacked bar chart with percentage of active and not active home links, cumulative statistics for each database is presented in the same manner as for SPARQL endpoint tests.



DATABASE: The European Library Dataset
DATABASE FULL NAME: The European Library Dataset
IS HOME LINK ACTIVE: YES
HOME LINK: http://www.theeuropeanlibrary.org/tel4/access/data/opendata/details
HAS SPARQL ENDPOINT: NO
DATE OF LAST TEST: 8. junij 2020
HTTP RESPONSE CODE: 200

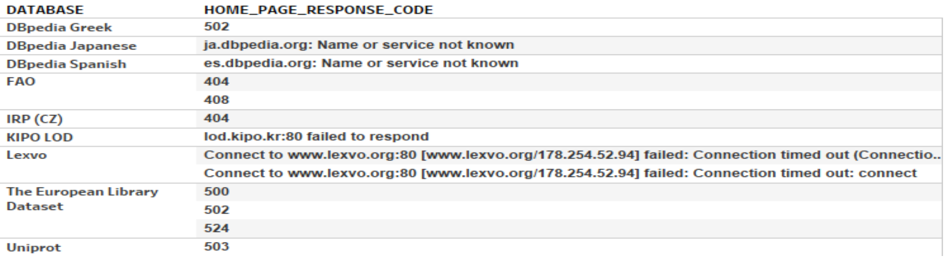
Dashboard shows whether database home link is active on last tes run. To be active means that HTTP request to home link returned response with 200 OK HTTP response code.
Graph **% OF SUCCESS/UNSUCCESS PINGS FOR DATABASE** show total percentage of pings when database was alive.
Graph **HOME LINK FOR ACTIVE FOR LAST 30 DAY**S shows when some of database's home link was inactive.
Click eihter on on one of squares on treemap to check databases statistics or click on database name on
% OF SUCCESS/UNSUCCESS PING graph or select database from dropdown menu.

Figure 15. *Example of dashboard with ping to database home links*

> On the iplod.io webpage by clicking on a database on a treemap or by selecting it from a chart with database cumulative statistics, only results for the selected database are shown in the time graph. Values for a single selected database are always 100% because a filtered chart shows statistics for all databases, while a filter is applied by selecting a single database, we only try to display dates when the selected database was not active..

The most common response codes in case of error were 500, 502, 503, 524 which mark Internal Server Error, Bad Gateway, Service Unavailable and Time out reached respectively which occur when the site is down or under maintenance while HTTP server is still running. Next were 404 and 408 meaning Not Found and Request Timeout respectively where Not Found means that resource for web page was not found on the url address, while Request Timeout is returned when the server tries to close the connection with the client. Both were returned while the server was running. Other errors were returned when the pages were not responding at all or server url was actually down.

Table 3. *List of HTTP codes and exceptions for home link ping test.*

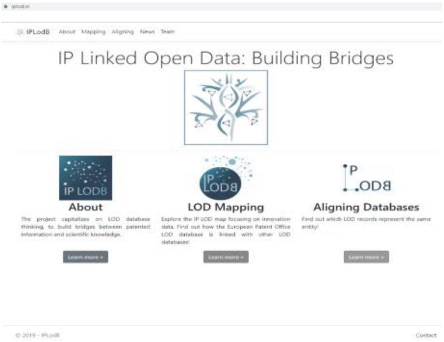


DATABASE	HOME_PAGE_RESPONSE_CODE
DBpedia Greek	502
DBpedia Japanese	ja.dbpedia.org: Name or service not known
DBpedia Spanish	es.dbpedia.org: Name or service not known
FAO	404
	408
IRP (CZ)	404
KIPO LOD	lod.kipo.kr:80 failed to respond
Lexvo	Connect to www.lexvo.org:80 [www.lexvo.org/178.254.52.94] failed: Connection timed out (Connectio..
	Connect to www.lexvo.org:80 [www.lexvo.org/178.254.52.94] failed: Connection timed out: connect
The European Library Dataset	500
	502
	524
Uniprot	503

Publishing results on Tableau servers and the iplod.io webpage

Tableau Desktop application enables users to design interactive visualizations for business intelligence with different chart types that can be grouped in dashboards with different control over visualizations. Controls could be applied in form of dropdown select menus, slider bars or clicks on graphical objects or legends. With such functionality relatively quick and easy development of interactive visualizations is available, without prior knowledge of web development, and with the option of integration of various data types, from excel and csv files, SQL and NoSQL databases, to web data storage platforms like Google Drive and Dropbox. Users can publish developed visualizations to Tableau Server where they become available as standard html pages and can be other users. Tableau Server is available as a separate product and is more suitable for larger enterprises, while regular users can publish content either on Tableau Online or Tableau Public services, which are Tableau Servers running on cloud.

The webpage containing the map (as well as the results of database alignment) is available at http://www.iplod.io/. It runs on Server info: iPLESK Windows M 16GB with SSL Certificates server with RapidSSL upgrade and an IP address – linked domain.



IP Linked Open Data: Building Bridges

About

The project capitalizes on LOD database thinking to build bridges between patented information and scientific knowledge.

LOD Mapping

Explore the IP LOD map focusing on innovation data. Find out how the European Patent Office LOD database is linked with other LOD databases!

Aligning Databases

Find out which LOD records represent the same entity!

Figure 16. *IP LODB landing page*

Browsability of the LOD Data: Constructing Two Auxiliary Tools

To be able to connect the datasets, we need to first understand the variables through which they can be connected. Although some other authors (e.g. Hassan et al., 2018) have mentioned using distinct tools such as the LIMES tool, we decided for our own approach. We build on our understanding of LOD databases contents, including their variables, their attributes (e.g. number of observations, min, max etc.) and the links between them (ontologies). For this reason, we constructed the LODStatTool.

Upon research about data extraction tools for LOD data from several articles we found out that most of the tools do not exist anymore so we decided to build a tool for our own needs. LODStatTool is a simple command line tool for analyzing RDF data implemented in C# .NET Core. Given an RDF input file the tool is capable of extracting predicates and upon specifying specific data types for given predicates it can generate simple statistics on the input data.

The statistics statistics generation currently supports the following data types: number, name, url, and date. The available statistics somewhat differentiates between different data types, but e.g. for number these are: Number of all elements processed, Number of unique elements, Minimal value, Maximal value, Mean value, Standard deviation.

For more in-depth handling of data, we also constructed the LODManager tool. The tool was initially used for the disambiguation and alignment of EP LOD and SN SciGraph (but is also more widely applicable. The LODManager functionality includes: downloading files from given URLs; reading files in different formats, e.g. N-Triples, JSON; storing data and data analysis. The LODManager is used also during the EP LOD-SN SciGraph alignment efforts.

The tools have have over 60 000 lines of code. The LODStat is a simpler tool, and is already available at *https://gitlab.com/IPLOD/lodstat*. The LODManager will be released at a later date.

References

1. Assaf, A., Troncy, R. and Senart, A. (2015). What's up LOD Cloud?. In European Semantic Web Conference (pp. 247-254). Springer, Cham. Available at: https://www.researchgate.net/profile/Ahmad_Assaf3/publication/286779461_What%27s_up_LOD_Cloud_Observing_The_State_of_Linked_Open_Data_Cloud_Metadata/links/566dd4cf08ae430ab5001d3c/Whats-up-LOD-Cloud-Observing-The-State-of-Linked-Open-Data-Cloud-Metadata.pdf
2. Auer, S. (2014). Introduction to LOD2. Linked Open Data - Creating Knowledge Out of Interlinked Data Results of the LOD2 Project (eds.: Auer, Soeren, Bryl, Volha and Tramp, Sebastian). Cham, Heidelberg, New York, Dordrecht, London: Springer, pp. 1-20.
3. Berners-Lee, T. (2006). Linked Data. A personal view note. Available at: https://www.w3.org/DesignIssues/LinkedData.html.
4. Belleau, François, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. "Bio2RDF: towards a mashup to build bioinformatics knowledge systems." Journal of biomedical informatics 41, no. 5 (2008): 706-716.
5. Bryl, V., Bizer, C., Isele, R., Verlic, M., Hong, S. G., Jang, S., ... and Choi, K. S. (2014). Interlinking and knowledge fusion. In Linked Open Data--Creating Knowledge Out of Interlinked Data. Cham, Heidelberg, New York, Dordrecht, London: Springer, pp. 70-89.
6. Dumontier, M., Callahan, A., Cruz-Toledo, J., Ansell, P., Emonet, V., Belleau, F. and Droit, A., 2014, October. Bio2RDF release 3: a larger connected network of linked data for the life sciences. In Proceedings of the 2014 International Conference on Posters & Demonstrations Track (Vol. 1272, pp. 401-404). Available at: http://ceur-ws.org/Vol-1272/paper_121.pdf.
7. Dividino, R.Q., Gottron, T., Scherp, A. and Gröner, G., 2014, From Changes to Dynamics: Dynamics Analysis of Linked Open Data Sources. In PROFILES@ ESWC. Available at: http://ceur-ws.org/Vol-1151/paper4.pdf?
8. European Patent Office. (2019). Linked open EP data. Retrieved 20 June 2019, from https://www.epo.org/searching-for-patents/data/linked-open-data.html#tab-1
9. Eurostat. (2019). Linked Open Data - Eurostat. Retrieved 20 June 2019, from https://ec.europa.eu/eurostat/web/nuts/linked-open-data
10. Hassan, M. M., Zaveri, A., and Lehmann, J. (2018). A linked open data representation of patents registered in the US from 2005–2017. Nature, Scientific data, 5, DOI: 10.1038/sdata.2018.279.
11. Hausenblas, M. and Karnstedt, M. (2010). Understanding Linked Open Data as a Web-Scale Database. Second International Conference on Advances in Databases, Knowledge, and Data Applications, 2010.
12. Hu, W., Qiu, H., Huang, J. and Dumontier, M., 2017. BioSearch: a semantic search engine for Bio2RDF. Database, 2017, 1-13. DOI: 10.1093/database/bax059.
13. Johnson, A.R., Hafner, A., Lužar, B., Modic, D., Rožac, B. and Vučković, M. (2019). Intellectual Property Linked Open Data: Building Bridges (IP LodB) Towards Developing Small Business Informatics. *Paper for the ITIS 2019 conference*, Slovenia November 7-9.
14. Khusro, S., J., F., Syed Rahman Mashwani S. and Iftikhar, A. (2014). Linked Open Data: Towards the Realization of Semantic Web-A Review. Indian Journal of Science and Technology, 7(6): 745-764.
15. Kracker, M. (2017). New PI product: Linked open EP Data. Presentation at IP Statistics for Decision Makers conference 2017, 14-15 November 2017. Authors' archive.
16. Lehmann, J., I., Robert, J., Max, J., Jentzsch, A. and Kontokostas, D. (2012). DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web, 1: 1–5.
17. Papadaki, M. E., Papadakos, P., Mountantonakis, M., & Tzitzikas, Y. (2018, March). An Interactive 3D Visualization for the LOD Cloud. In EDBT/ICDT Workshops (pp. 100-103).
18. Polleres, A., Kamdar, M. R., Fernández, J. D., Tudorache, T., & Musen, M. A. (2020). A more decentralized vision for linked data. Semantic Web, (Preprint), 1-13.
19. Springer Nature. (2019). SciGraph | For Researchers | Springer Nature. Retrieved 20 June 2019, from https://www.springernature.com/gp/researchers/scigraph
20. The Linked Open Data Cloud. (2019). Retrieved 20 June 2019, from https://lod-cloud.net/#
21. Umbrich, J., Hausenblas, M., Hogan, A., Polleres, A., & Decker, S. (2010). Towards dataset dynamics: Change frequency of linked open data sources.
22. World Bank. (2019). About us | Data. Retrieved 20 June 2019, from https://data.worldbank.org/about

Appendix 2

IP LodB project
METHODOLOGY REPORT
(Deliverable 3.1)

2020

Table of Contents

This report has been prepared by the members of the IPLOD group: Dolores Modic, Alan Johnson, Borut Lužar, Ana Hafner, Borut Rožac, Miha Vučkovič and Einar Rasmussen for the European Patent Office (EPO) as part of the reporting inside the IP linked open data: building bridges (IP LodB) project. The IP LodB project is an EPO Academic Research Programme project (https://www.epo.org/learning/materials/academic-research-programme.html).

Please cite as: Modic, D., Johnson, A. Lužar, B., Hafner A., Rožac B., Vučkovič M. And Rasmussen E. (2020). IP LodB project Methodology report for EPO Academic Research Programme. Available at: https://www.epo.org/learning/materials/academic-research-programme/research-project-grants.html.

The report is also available at: https://www.researchgate.net/publication/343690785_Intellectual_Property_Linked_Open_Data_Building_Bridges_IP_LodB_project_Methodology_Report.

Executive Summary

The “Intellectual Property Linked Open Data Building Bridges” (IP LodB) project, with a multidisciplinary team and funding from the European Patent Office Academic Research Programme (ARP) funding, develops connections from bibliographic data for patent documents to other linked open data (LOD) sources available on the world wide web. IPLodB is envisaged to become an important source of linkability between linked open data (LOD) sources on basic research, knowledge production, innovation and technology management, because it helps to map pathways from science and engineering research activities to codified knowledge.

In many contexts, technological progress and innovation is essential to national and regional economic growth and well-being. Archival data from secondary sources have well-known advantages and disadvantages compared to primary data, e.g., surveys and case studies, for use in the social and behavioral sciences (Wenzel & Van Quaquebeke, 2018). Put simply, the former have lower bias and the latter have lower variance. Analyses of patent and bibliographic data can provide a dynamic picture of regional and organizational competences and quantify benefits from public policy interventions and collaborations between organizations that are useful to public policy makers, business leaders, and social and behavioral science researchers.

The European Patent Office (EPO) LOD source is the principal focus of our project. Our overall objective was to evaluate its linkability to other LOD sources and recommend strategies for improvement. The IPLodB has two deliverables: a “Map” (IPLodB Map) and “Database” (IPLodB Database). The outcome of the project is an **Enhanced LOD Space** in the area of intellectual property.

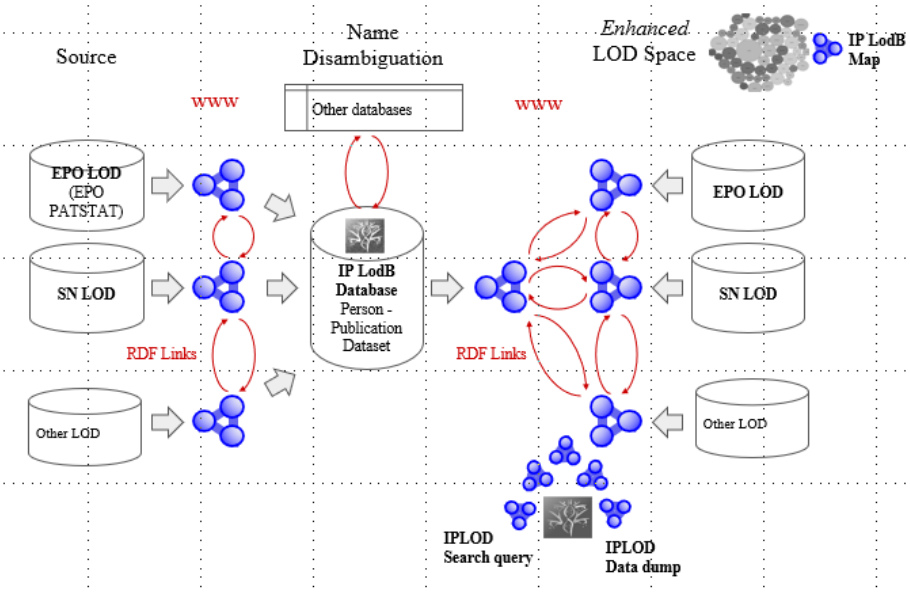


Figure 1. Enhanced Linked open data space.

The IP LodB project focuses on interlinking the EPO LOD and SN LOD database, using some other identified LOD databases (e.g. GRID) to build further links in the LOD space. Other databases. not currently in the LOD format, are also used to enrich the IP LodB Space to begin building an LOD innovation subcloud around the EPO LOD database as the hub. At the same time linkeability across the LOD innovation data space increases and the results can cross-fertilize the existing LOD datasets, if adopting the suggested links. The results are fed back to the world wide web, both via an operational queryable search option, and through a data dump. Based on Khusro and colleagues (2014, p. 748).

Today, several databases from reputable sources are available in the Linked open data (LOD) format, where linked data *inter alia* denotes that each record has its own uniform resource identifier (URI).

The four principles of linked data (Berners-Lee, 2006) :
1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs so that they can discover more things.

The idea behind this is that linked data would be one more step towards the web transforming from the web of documents to web of data. This project is an early attempt to enrich the LOD landscape (in the field of intellectual property), which is still not close to its vision of the data on the whole web being able to be searched as one big database.

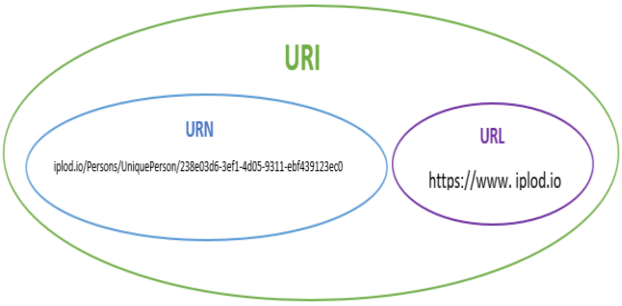

URI
URN
iplod.io/Persons/UniquePerson/238e03d6-3ef1-4d05-9311-ebf439123ec0
URL
https://www. iplod.io

Figure ES1: Uniform resource identifier as the cornerstone of linked data

Uniform resource identifier (URI) has two components; the uniform resource name (URN), which should be denoting a unique real-life object; and the uniform resource locator, which shows the path to this data and how to retrieve it (note that according to Berners-Lee (2006) definition, linked data uses the standard http protocol. Linked data should thus be carefully handled (i.e. disambiguated), classified, matched, and interlinked – taking into account the prevailing standards.

The IPLodB Map is a qualitative evaluation of LOD sources, e.g. Springer Nature (SN) LOD, GRID, that are potentially linkable to EPO LOD either now or in the future. The connected work also entails checking their availability via a developed pinging protocol. We use the realizations acquired during constructing the Map (presented in 2.1) to enhance our work on the IP LodB Database.

The IP LodB Database is a quantitative evaluation of linkability between EPO LOD and SN LOD (presented in this document), which includes an intervention to increase the uniqueness of both sources by introducing a *nascent* person identifier based on disambiguation procedures for name text strings, i.e., for inventors and authors, and other person attributes, e.g., collaboration, affiliation, location, and content. Specifically, the IP LodB database disambiguates the names of people and organizations associated with publications, i.e., patent documents, journal articles (books, book chapters). The disambiguation is accomplished using relational database management systems (RDBMS) harmonization techniques. We also evaluate the reliability and validity of various combinations of techniques using machine learning (ML) procedures.

Disambiguation should be highly prioritized in LOD sources (since they provide URIs, ie. uniform resource identifiers), however especially for LOD data aggregators, which connect data across various LOD datasets, we have found little evidence this would indeed be the case. The IPLOD project however relies on a robust disambiguation method applied both within the databases as well as when aligning the databases in focus.

A golden dataset for disambiguation

A hand curated training (annotated and checked) set of 2,895 distinct persons before disambiguation, equaling 1,111,771 pairs of original entities that have been manually confirmed to represent a match.

We provide validation for our disambiguation method and show that our disambiguation method demonstrates a significant improvement over more naive approaches, both in terms of precision and recall.

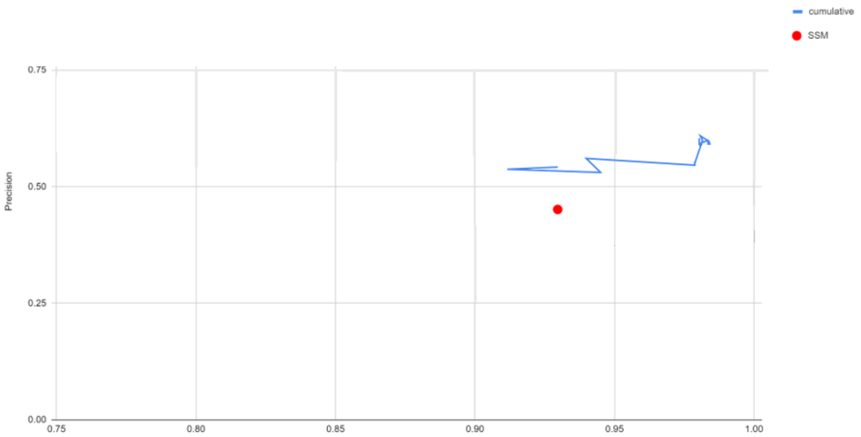


Figure ES2: Precision and recall – comparison to simple string match (SSM)

Higher values for both recall and precision reflect better matching. Above we show a very conservative approach when showing calculations; not including trivial cases (i.e. only including names that are likely to be matched. If we would use a more common method, than our rate for false negatives would be 1.93% and for false positives would be 1.98%.

When comparing this to the disambiguation fields in Patstat, i.e. HAN_ID and PSN_ID, which are optimized for legal persons (as inventors or applicants), the present disambiguation is able to detect considerably more matches (true positives) than the now used ones.

Our disambiguation effort allows us to significantly reduce the number of unique persons both within single databases as well as across both databases. These interconnected have an *iplod nascent person identifier* attributed.

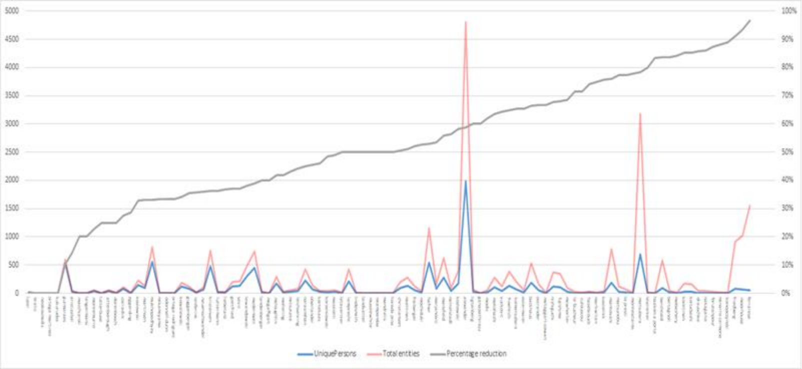
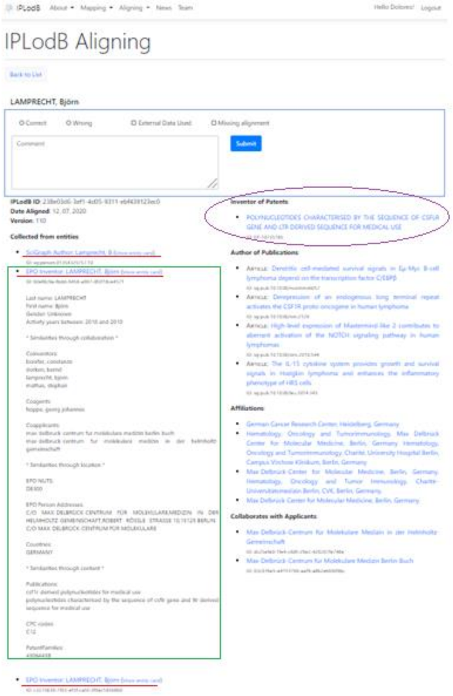


Figure ES3: Reduction in unique persons pre- and post-disambiguation (n=25958)

We are then able to diligently interconnect persons across both focal LOD databases, i.e. the EPO LOD and SN LOD. We thus consolidate all the available data on a specific real-life object (i.e. person) under a specific URI, which contains so-called entity card.

We interconnect approx. 1.9 % of all persons between EPO LOD and SN LOD and provide links between them.

The final part of our project was to publish the IP LodB dataset with increased linkeability between the EPO LOD and SN LOD sources back out onto the world wide web, which is done via a queryable search option on the iplod.io webpage that includes hyperlinks to especially EPO LOD and SN LOD datasets - thus building a bridge between the two databases. Since the human-readability of the LOD data is often a problem, we provide a user friendly opportunity to query an already interlinked set of data.


IPLodB About ▾ Mapping ▾ Aligning ▾ News Team
Hello Dolores! Logout
Back to List
○ Correct ○ Wrong □ External Data Used □ Missing alignment
Comment
Submit

IPLodB Aligning

LAMPRECHT, Björn

IPLodB ID 23be03d6-3ef1-4c05-9311-eb4438123ec0
Date Aligned 12. 07. 2020
Version 110

Collected from entities

- SciGraph Author Lamprecht, B (show entity card)
  [illegible]
- EPO Inventor LAMPRECHT, Björn (show entity card)
  [illegible]

  Last name: LAMPRECHT
  First name: Björn
  Gender: Unknown
  Activity years between: 2010 and 2013

  * Similarities through collaboration *

  Coinventors:
  bonifer, constanze
  dorken, bernd
  lamprecht, björn
  mathas, stephan

  Coagents:
  hoppe, georg johannes

  Coapplicants:
  max delbruck centrum fur molekulare medizin berlin buch
  max delbruck centrum fur molekulare medizin in der helmholtz gemeinschaft

  * Similarities through location *

  EPO NUTS:
  DE300

  EPO Person Addresses:
  C/O MAX DELBRÜCK CENTRUM FÜR MOLEKULARE MEDIZIN IN DER HELMHOLTZ GEMEINSCHAFT ROBERT RÖSSLE STRASSE 10 13125 BERLIN
  C/O MAX DELBRÜCK CENTRUM FÜR MOLEKULARE

  Countries:
  GERMANY

  * Similarities through content *

  Publications:
  csf1r derived polynucleotides for medical use
  polynucleotides characterised by the sequence of csfr gene and ltr derived sequence for medical use

  CPC codes:
  C12

  PatentFamilies:
  41094418
- EPO Inventor LAMPRECHT, Björn (show entity card)
  [illegible]

Inventor of Patents

- POLYNUCLEOTIDES CHARACTERISED BY THE SEQUENCE OF CSF1R GENE AND LTR DERIVED SEQUENCE FOR MEDICAL USE
  [illegible]

Author of Publications

- Article: Dendritic cell-mediated survival signals in Eμ-Myc B-cell lymphoma depend on the transcription factor C/EBPβ
  [illegible]
- Article: Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma
  [illegible]
- Article: High-level expression of Mastermind-like 2 contributes to aberrant activation of the NOTCH signaling pathway in human lymphomas
  [illegible]
- Article: The IL-15 cytokine system provides growth and survival signals in Hodgkin lymphoma and enhances the inflammatory phenotype of HRS cells
  [illegible]

Affiliations

- German Cancer Research Center, Heidelberg, Germany
- Hematology, Oncology and Tumorimmunology, Max Delbrück Center for Molecular Medicine, Berlin, Germany; Hematology, Oncology and Tumorimmunology, Charité, University Hospital Berlin, Campus Virchow Klinikum, Berlin, Germany
- Max Delbrück Center for Molecular Medicine, Berlin, Germany; Hematology, Oncology and Tumor Immunology, Charité-Universitätsmedizin Berlin, CVK, Berlin, Germany
- Max Delbrück Center for Molecular Medicine, Berlin, Germany

Collaborates with Applicants

- Max Delbrück Centrum für Molekulare Medizin in der Helmholtz-Gemeinschaft
  [illegible]
- Max Delbrück Centrum für Molekulare Medizin Berlin Buch
  [illegible]

Figure ES4: The user friendly opportunity to query results with links to original sources

The Figure displays the so-called entity card (green box) for a person record which was disambiguated from SN and LOD original (pre-disambiguation) persons (underlined in red). The full entity card, which is an enhanced person-publication record we provide and contains all of the fields for a certain record, is available to registered users. To non-registered users a more limited information on an entity is provided (right side of the Figure). Nonetheless, the enforced links to original data sources (especially EPO LOD, SN LOD and GRID) are provided (e.g. link in purple circle).

Furthermore, we provide a bulk download in RDF format for blocks of data, which are available from the iplod.io webpage.

<https://data.epo.org/linked-data/data/vc/D5890E976016F4B6990707BD1D049EB5>
<**http://www.w3.org/2002/07/owl#sameAs**>
<https://www.iplod.io/Persons/UniquePerson/3d546f57-7103-40ab-8b6e-012978829e89>

Figure ES5: The RDF format triple (EPO and LOD)

A triple has the format subject-predicate-object and thus contains a simple statement. Here the standard predicate to connect two URIs and designate them as representing the same real-life object is used (the "owl#sameAs"). The triple connects a person from the EPO LOD to an iplod disambiguated and interconnected (with SN LOD) person.

The IPLOD Database allows from the more common analysis within database and the evolution of patenting and publishing behaviors separately, also to cross the two, via persons that both patent and publish. Hence, we can observe also variations in their patenting and publishing behaviours. Furthermore, the LOD format allows the original providers of data to add data, which can be additionally discovered when following the links from a data bridge, like the IPLOD.

LOD format databases can facilitate future applied research into the innovation process by both people and machines.

Recommendations

We provide several recommendations. Firstly, for the EPO LOD database itself and its usability, we suggest the following:

- Add to the data catalogue of the EPO LOD database, by providing also the appl_id, person_id and abstracts
- Providing EPO LOD data in back and front file formats, similar to Patstat data.
- Further changes in server to support the SPARQL, which is now not sufficient to support more complex querying.
- Providing information on the EPO LOD-DBPedia test done (Kracker, 2017), to support further work that can be done on the interconnection between DBPedia and EPO LOD or Wikidata and EPO LOD

Secondly, in terms of higher interconnectivity within the EPO LOD database and allowing more discoverability of other data from the EPO LOD database:

- Using «same as» links to link records denoting same real-world objects (especially inventors initially) within EPO LOD database. These can be grabbed through bulk download at iplod.io webpage.
- Using «same as» links to IP LodB database to allow users to discover more information – this way the same as links lead to the iplod.io database, which allows

the users to discover more information in accordance with the linked open data principles.

- Adopting the disambiguated persons IDs for natural persons also in other EPO products, e.g. Patstat, with a possible further precision and recall calibration.

Further planned work connected to IPLOD database

Honing the algorithm for natural persons and running it on all blocks and providing links, i.e. providing bi-monthly new RDF packages with links. For EPO LOD and SN LOD alignment, we will work on further improvement of the RAKE keywords method and exploring alternative opportunities. We are also publishing the article on method and on applying this to the area of Circular economy.

Furthermore, we will be shifting our focus to disambiguating the legal entities and establishing additional links between EPO LOD and other databases in terms of included organizations (applicants), subsequently running the algorithm on all blocks and providing links.

Introduction

Research into invention, innovation policy, and technology management can benefit from accurate *person* identifiers for individuals engaged in research activity that would help keep track of individuals' scientific publications and, by extension, overall progress in various fields of science and engineering. To this end, the European Patent Office (EPO) and Springer Nature (SN) are leaders in publishing bibliometric data on research and innovation in a linked open data (LOD) format. Both EPO and SN, similar to other patent offices and publishers around the world, provide lists of unique *publications*, i.e., patent documents and journal articles (books, book chapters),[1] but the lists of *people* they provide are not unique, i.e., inventors and authors. Specifically, a person who collaborated on two or more publications is likely to appear on the list under several different identifiers, i.e., ostensibly as several different people. In recent years, scientometric scholars have developed some ingenious techniques to add a *nascent* person identifier to bibliographic data sources using text strings for family name and given name fields, combined with text strings from several other fields, e.g., collaborator, affiliation, location, and content, which taken together are indicators of an underlying nascent person identifier. Our project builds on previous name disambiguation research to improve the entity resolution both within and between the EPO LOD and SN LOD bibliographic data sources using RDBMS harmonization techniques (Li et al., 2014; Morrison et al., 2017; Torvik & Smalheiser, 2009).

Previous research on patent documents suggests that at least 10% of bibliographic records suffer from problems with ambiguity and that simplistic procedures to accommodate the anomalies produce material differences to analyses of findings from data (Raffo & Lhuillery, 2009). Similar work on journal articles suggest the situation there is considerably worse because given names are often reduced to initials and author addresses are systematically missing (Torvik & Smalheiser, 2009). Accurate disambiguation of people presents at least three related challenges. *First*, without a unique list for people, how can we know who these five text strings refer to: 'Donal Duck', 'Donald Duck', 'Daffy Duck', and two observations with only initials for given names, 'D.H.T. Duck' and 'D.F. Duck'? Donal and Donald are affiliated to The Walt Disney Company and Daffy is affiliated to Warner Bros., further all five are researchers on water installations or water treatment, but what about the two observations with just initials, especially where affiliation and content are ambiguous. How can we infer whether these records refer to the same or different people without manual investigation. *Second*, how can we repeat such an investigation for several million records across two or more bibliographic data sources. And, *third*, even if we come up with some scoring procedure to ameliorate ambiguity in the source data, despite validation and augmentation from secondary sources, the brute force approach to compare every record with every other record is not feasible with tens of millions of records involved. An accurate and automated procedure to create a nascent person identifier from disambiguated person

[1] The bibliographic data from SN LOD include journal articles, books, book chapters but we sometimes refer to all three together as 'journal articles'.

attributes, e.g., name, collaborator, affiliation, location, content, from the EPO and SN LOD sources requires a careful iterative RDBMS design and, even then, intensive computation is needed to execute the procedures.

Our approach has contributed both conceptual clarity and technical innovation to the name disambiguation literature. On the conceptual side, we have reviewed the scientific concepts and operational definitions used in previous research on name disambiguation for bibliographic data. On the technical side, we have proposed a creative application of conventional RDBMS technology to add value to leading edge LOD technology, which is pivotal to democratize the availability of data. Finally, we release the data back to the wide world web in two formats that bridge the datasets; one is through a user interface and with search query, including a wiki-type dialogue box for users to propose edits, where the links lead back to LOD source data (not only EPO LOD and SN LOD, but also GRID) and the other in the format of RDF links in line with our data publishing module, for already processed blocks.

To allow for the validation of the model and to insure quality, records in several versions have undergone manual checking, but for the current version (110) we also built a golden dataset (i.e., a training set), which is a hand curated (annotated and checked) set of 2,895 distinct persons before disambiguation, or in other words 1,111,771 pairs of original entities that have been manually confirmed to represent a match. The results of our algorithm on a smaller sample of data are very promising and outperform simple algorithms that rely on trivial data, such as the naive one, simply comparing matches on family and given names. Furthermore, we observe significant rates in reduction of unique persons within surnames, compared to original data disambiguation. Our test is done on records connected to 100 randomized surnames, within which were 37,832 original (i.e. pre-disambiguation) persons. Lastly, our initial calculations show we are able to interconnect about 1.9 % of persons, i.e. these are our unique persons, which are derived from both EPO LOD and SN LOD pre-disambiguated persons.

Initial tests show that the problem of analyzing the whole dataset is computationally feasible in practice; namely, it will take a couple of months to evaluate all blocks in the first iteration. Once the complete IP LodB dataset is established, the disambiguation within other sources should be much simpler, since data about a number of entities will already be aggregated in our database and so the entities from new sources can immediately be compared against IP LodB entities.

Background - Name Disambiguation and Scientometrics

The demand for scientific and technology indicators is growing in the face of technical challenges from resource sustainability, circular economy, and climate change, among other things. Demand comes from both public policy makers and business leaders who want to evaluate innovation output from both public research organizations (PROs) and the private sector. At the moment, demand is acute in Europe because public policy makers want reliable data on university inventions, publications, and technology transfer, including valid international comparisons. Business leaders and professional investors are also interested in science and technology indicators to monitor potential opportunities and competitive threats.

In principle, LOD protocols make data publicly available and if interlinked they can address the problem of integrating heterogeneous data, which is an inherent problem of big data (Sleeman et al., 2015). The final vision for the LOD technology is that all the data available on the world-wide-web will be treated and researched as one database with the aim to share and reuse existing data (Khusro et al., 2014), while propagating a machine-readable format. However, while the amount of available LOD has grown significantly over the last years, most data sources are still not sufficiently interlinked. The difficulties in identifying entities in same or different data sources that describe the same real-world object persist and the vocabularies and ontologies remain not unified. Many LOD sources are difficult to discover with their availability and accessibility remaining unstable. Most challenges stem from the nature of LOD: on one hand, it is open and distributed but, on the other hand, it is plagued by low connectivity. Its ontology-based approach allows machines to understand data, however the tools allowing humans to browse through the data are lacking. The semantic web 'brochure' promises rich and self-describing relations from efficient (multi-source) database utilizations (Bryl et al., 2014) but, typically, the promise is not fulfilled for emerging data dimensions, i.e., RDF triples.

Our primary bibliographic data sources are the EPO LOD (EPO, 2019) and the SN LOD (SN, 2019), respectively. The EPO LOD contains information on **3.4 million** unique patent documents for 90 countries, going back to the 1830s for some patent authorities.[2,3] The SN LOD contains analogous information on **9.7 million** unique journal article (books, book chapter) records, also going back to the 1830s.[4,5] We have retrieved data about **13.2 million** publications for the combined data sources, which is a considerable increase in scope compared to previous work, e.g., patent documents from 1977 up to the present (Li et al., 2014; Pezzoni et al., 2014), journal articles from Medline (Torvik & Smalheiser, 2009). Both bibliographic data sources are updated regularly, the EPO LOD (available in Ntriples format)

[2] The EPO LOD contains data for 3,449,733 unique patent documents that include some inventor name data.

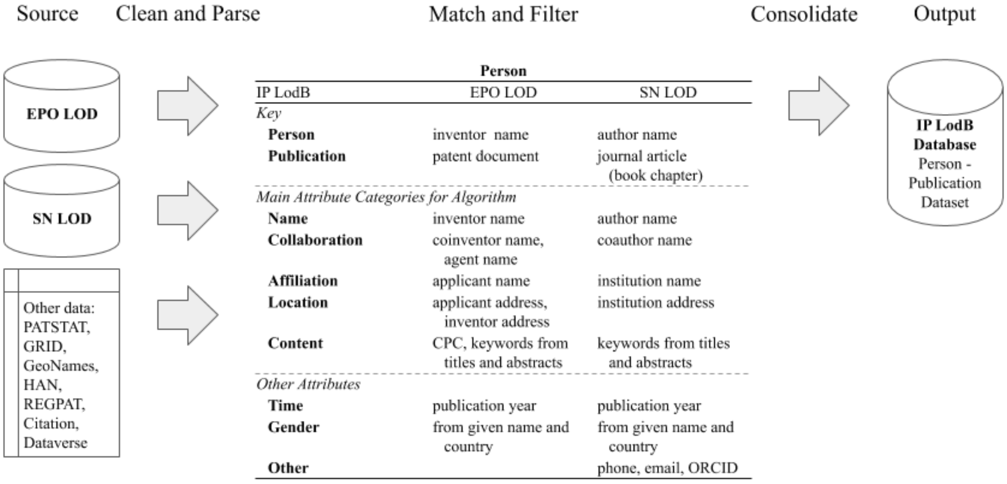
[3] Note that inside the EPO LOD there are 23,391,587 patent documents with URIs some of which are duplicates to the core patent document and some cover the so-called non-EP documents, where the information is limited to basic patent identification information. Thus, some have only limited data, which does not include either an applicant or an inventor. Nonetheless, they contain the application number as well as have a unique URI (see as e.g. https://data.epo.org/linked-data/doc/application/AL/38172011P). Last category are sheer duplicates (see e.g. https://data.epo.org/linked-data/doc/application/AT/80362010U and https://data.epo.org/linked-data/doc/application/AT/803610U), which were already noted by the IPLOD group to EPO and will be corrected in future versions of the EPO LOD by the EPO.

[4] The SN LOD contains data for 9,781,053 publications that include some author name data, where we can add value through name disambiguation procedures. Publications, which include some name data, broken down by type include: unique journal articles: 5,606,284; books, 133,270; book chapters, 4,041,499.

[5] In addition, the SN LOD contains data on another 2,898,266 publications with *missing* author name data, where the probability of name disambiguation is small. These publications contain other material that research reports, e.g., tables of contents, indexes, editorials, letters to the editor, calls for papers and other advertisements, so it is not a surprise to find some documents missing author name data. Publications with missing name data broken down by type include: unique journal articles, 2,440,515; books, 124,704; book chapters, 333,047. In total, SN LOD contains data for 12,679,319 publications with 77.1% usable for name disambiguation, which includes: unique journal articles, 8,046,799 (69.7% usable); books, 257,974 (51.7% usable); book chapters, 4,374,546 (97.4% usable).

is updated weekly but the SN LOD (available in json format) less frequently, at the moment, most download files are dated February 2019.

Secondary sources are used to add attributes to or validate attributes from LOD sources. Organization name strings come with various spellings and subsidiary details, e.g., International Business Machines (United States), IBM (Germany), and to help resolve these anomalies source fields can be validated using the GRID (Digital Science, 2020) database of research organizations or the harmonized applicant name (HAN) database (Magerman et al., 2006) of alternate spellings for patent assignees. Inventor addresses are not published in the EPO LOD, so we retrieved them from PATSTAT (EPO, 2020). Address strings for both individuals and organizations also come with various degrees of accuracy and resolution, e.g., building number, street name, post code, state, country. Anomalies in address data can also be reconciled using the GeoNames (GeoNames, 2020) database of place names or the OECD REGPAT database (Maraut et al., 2008) of patents by region. In addition, there are higher resolution geographic location solutions available when either building number and street name (Morrison et al., 2017) or post code are available.


Source
Clean and Parse
Match and Filter
Consolidate
Output
EPO LOD
SN LOD
Other data: PATSTAT, GRID, GeoNames, HAN, REGPAT, Citation, Dataverse
IP LodB Database Person - Publication Dataset

Person		
IP LodB	EPO LOD	SN LOD
Key		
Person	inventor name	author name
Publication	patent document	journal article (book chapter)
Main Attribute Categories for Algorithm		
Name	inventor name	author name
Collaboration	coinventor name, agent name	coauthor name
Affiliation	applicant name	institution name
Location	applicant address, inventor address	institution address
Content	CPC, keywords from titles and abstracts	keywords from titles and abstracts
Other Attributes		
Time	publication year	publication year
Gender	from given name and country	from given name and country
Other		phone, email, ORCID

Figure 2. Data preparation process for name disambiguation.

The Figure illustrates the disambiguation process from name text strings to nascent person identifier for EPO and SN LOD bibliographic data sources. The IP LodB database is created through three stages, i.e., clean and parse, match, and filter, using RDBMS attributes, keys, and tables. In addition to name, person attributes are organised into four main categories, i.e., collaboration, affiliation, location, and content, for use in our algorithm. Based on Li and colleagues (2014, p. 943).

Figure 2 provides an overview of existing LOD sources, and our data preparation process and ultimate objective, to provide a bibliographic LOD source with high resolution nascent identifiers for people, i.e., the IP LodB database, using disambiguation techniques for name text strings and other person attributes. Figure 2 illustrates how we create the person table, which is the input data for the disambiguation process, from preliminary tables, where the source data is cleaned and parsed. Name disambiguation, also known as entity resolution (Idrissou et al., 2020), is an important operation within most large scale bibliographic data projects (Pezzoni et al., 2014). It consists in assigning nascent person identifiers, i.e., inventor or author, to a set of publication records, i.e., patent documents and journal articles, that are based on the same or very similar text strings for names and some combination of other attributes for a person, including their name, collaborators, affiliations, location, and content. We provide an overview of our labels for attributes, keys, and tables, in Figure 2 because consistent labels and well organized tables are powerful weapons against chaos in any data science project (Bryan, 2018).

Name disambiguation can be done in several different ways (Smalheiser & Torvik, 2009) but the two main approaches are: 1) hand-checking and collation by individuals and groups and 2) data science approaches using some combination of RDBMS procedures and machine learning (ML). Specifically, ML applications use supervised learning to develop a model with a training sample and, then, unsupervised learning to apply the model to the rest of the data (Pezzoni et al., 2014). Our approach leans towards RDBMS harmonization procedures, which emphasise iterative rule governed procedures to clean and parse, match, and filter, source data into name disambiguated data with a nascent person identifier. However, we also undertake a considerable amount of hand-checking and collation to develop a substantial training sample. Following Raffo and Lhuillery (2009), name disambiguation techniques can be described in three stages:

1) The *clean and parse* stage converts text strings from source data to input fields for name disambiguation, e.g., family name, given name, organization name, address, to a standard character set, which can accommodate minor spelling errors using token and lexical distance techniques.
2) The principal objective of the *match* stage is to reduce the number of false negatives in the LOD source. The match stage finds pairs of person-publication records that refer to the same person. If two people have the same or very similar names and also the same or very similar values on some combination of other attributes, i.e., collaborator, affiliation, location, content, then the two records probably refer to the same person. However, while we strive to reduce false positives, i.e., lumping, to a minimum, they sometimes occur.
3) The principal objective of the *filter* stage is to reduce the number of false positives from the match stage. The filter stage finds and rejects false positive matches for person-publication records that do not refer to the same person. Thus we use alternative attribute combinations and blocking rules, to check the robustness of matches from previous rounds of agglomerative clustering.

There is an emerging consensus that matched pairs of records should not rely solely on a simple string match (SSM) on names because they inflate both false positive results, i.e., lumping, by ignoring numerous different people with the popular names, e.g., Smith, John, and false negative results, i.e., splitting, minor mis-spellings, e.g., Duck, Donal, and occasional inversions of family and given names, i.e., Donald, Duck. Popular names can be disentangled using other attributes, e.g., collaboration, affiliation, location, and content. Minor misspelling and occasional inversions can be accommodated by decomposing text strings into component tokens of two characters or more and reassembling them using lexical distance, e.g., 2-gram, Levenstein distance (Pezzoni et al., 2014). Filters can identify and reject false positives, i.e., lumping, matches between records using alternative parameterizations of attributes to check robustness (Raffo & Lhuillery, 2009).

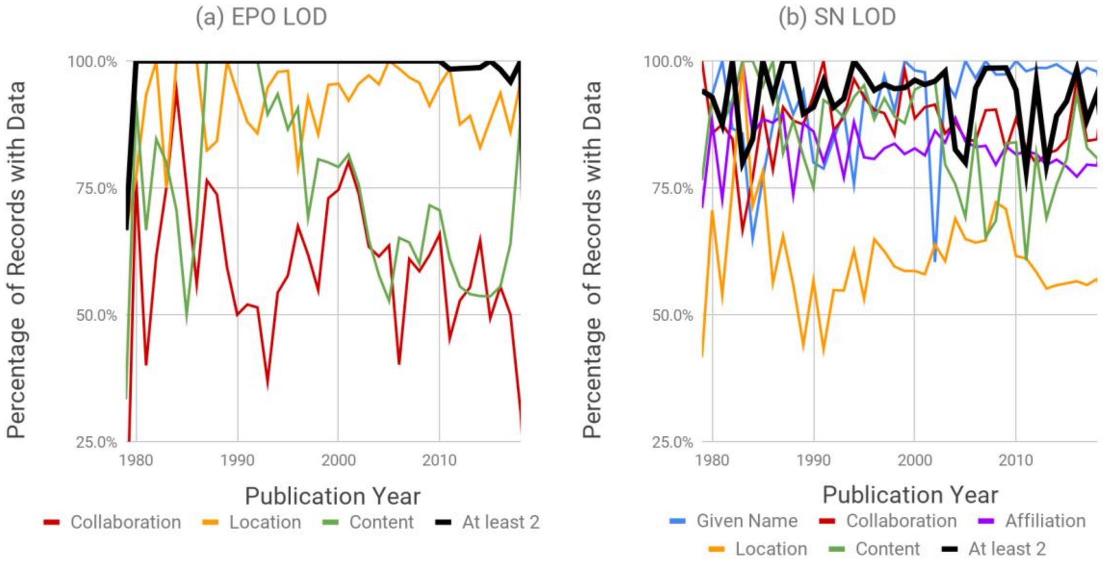


Figure 3. Available data on generic attribute categories by publication year and bibliographic data source.

The Figure shows the percentage of records with attribute data available by publication year and LOD source for our hand disambiguated dataset of 4,760 publications. Panel (a) for EPO LOD and panel (b) for SN LOD. The attributes include: given name, collaboration, affiliation, location (address), content (titles /abstracts).[6] The Figure does not show given name or affiliation for EPO because they are 100% throughout. Almost every publication has data for at least one additional attribute, i.e., collaboration, affiliation, location, content: EPO, 99.6%; SN, 95.0%. Indeed, most publications have data for at least two additional attributes: EPO, 98.9%; SN, 92.9%. The figure shows input data from 4,760 publications between 1979 and 2018: 1,944 from EPO and 2,816 from SN. Based on Torvik & Smalheiser (2009, p. 23).

Figure 3 illustrates the availability of data on important person attributes by publication year and LOD source after we have validated and augmented person-publication records from primary LOD sources with all available data from secondary sources. Representing author given names with only initials is a major disadvantage for name disambiguation of journal articles (books, book chapters), compared to patent documents that typically offer names spelt out in full (Li et al., 2014). Another difficulty for large scale name disambiguation is missing data on other attributes, i.e., collaboration, affiliation, location (address), content (titles /abstracts). There are no missing data for given name in the EPO LOD. However, Figure 3(b) suggests only 91.8% of person-publication records have some data on given names and previous literature suggests that many of these will be just initials. However, our experience working with these data suggests it is indeed possible to disambiguate records with missing data on given name, provided data on other attributes are available, i.e., collaboration, affiliation, location, content. So, what do we know about data available on other attributes. Almost every publication has some data available on at least one other attribute: EPO, 99.6%; SN, 95.0%.[7] Figure 3(a) suggests that most patent documents

[6] There is no address data in EPO LOD, so we retrieved them from PATSTAT.

[7] The percentage calculations are averages over the 39 years between 1979 and 2018. The percentages of available attribute data sometimes use person-publication records in the denominator and sometimes just publications. The relevant denominator is specified in the text.

have some data on least two other attributes, 98.9%. Figure 3(b) suggests that many journal articles also have data on least two other attributes, 92.9%. Taken together the Figure shows that, while name disambiguation of bibliographic data for journal articles has challenges compared to patent documents, the problem can be made tractable using names and four other attribute categories in a systematic name disambiguation procedure.

Digging a little deeper into the available data on names and other attributes, shown in Figure 3, sets the stage for the clean and parse, macte, and filter steps that we describe in the pages that follow. Collaboration between co-inventors or co-authors, is probably the second most accurate disambiguation attribute after name. Figure 3(a) suggests that many inventor-patent records have single inventors because only 58.9% of records have collaboration data, at least in this sample. However, Figure 3(b) suggests that most author-article records have two or more authors because 87.5% of journal articles have some collaboration data. Affiliation of inventors to applicants or authors to institutions is probably the second most accurate disambiguation attribute. All inventor-patent records have some affiliation data. Figure 3(b) also suggests that most author-article records have affiliation data for two or more authors because 83.0% of records have some affiliation data. Location data is not provided in the EPO LOD but we retrieved it from PATSTAT. Author address data is absent in the SN LOD but institution address data is often available, especially for the first (or corresponding) author.[8] Figure 3(a) suggests that most inventor-patent records have some location data because 91.8% have some location data, i.e., including co-inventors. However, Figure 3(b) suggests that many author-article records are missing location data, i.e., at least for second and third authors, because only 61.2% of records have location data. Content data is provided using the CPC taxonomy for the EPO LOD, but to integrate content across LOD sources we have created our own keywords based on publication titles and abstracts, by adapting a leading edge machine learning procedure (Rose et al., 2010). Unfortunately, Figure 3(a) suggests that many patent documents are missing content data because only 74.1% of them have abstracts. In comparison, journal articles typically have more content data because 85.3% of them have abstracts. Many of the issues raised here are returned to again and again through the theory, method, and results sections of the report.

Matching person-publication records in bibliographic data requires algorithms to process text strings containing person name fields and other attributes in non-trivial ways. Until recently, the black-box inside name disambiguation algorithms was kept relatively secret and sometimes protected, e.g., Massacrator©, but recently scholars appear more willing to discuss the details of different techniques. Moving towards ‘best’ procedures, Figure 4 suggests three things to consider when applying the clean and parse, match, and filter stages. *First*, cleaning and parsing by itself can *only* improve the recall rate, i.e., reduce false negatives or splitting. *Second*, filtering can improve the precision rate, i.e., reduce false positives or lumping. *Third*, the extent of the increase to the precision rate from filtering is contingent on an effective cleaning and parsing technique (Raffo & Lhuillery, 2009).

[8] In SN LOD, some more recent records contain email, phone numbers, and ORCID fields, but typically only for the corresponding author and not in sufficient quantity to be that helpful but, nevertheless, this is encouraging for future work.

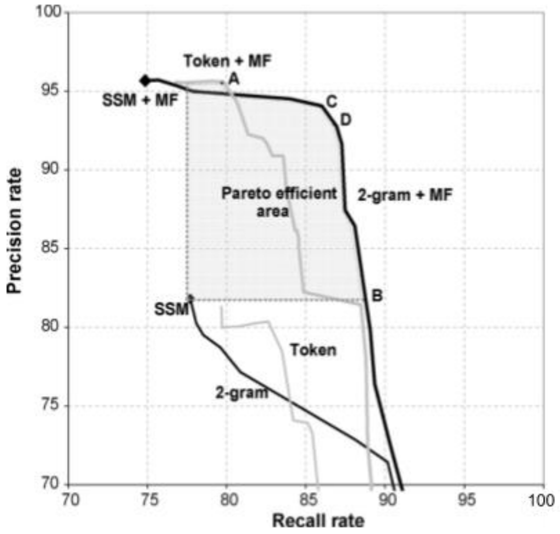


Figure 4. Precision and recall rates to evaluate reliability and validity.

The figure shows precision and recall points and curves for different parsing techniques and multiple filters (MF). Recall rate is *inversely* proportional to false negatives. Precision rate is *inversely* proportional to false positives. Higher values for both recall and precision reflect better matching.[9] The graph shows gains in both precision and recall rates compared to a simple string match (SSM) baseline.[10] Source Raffo and Lhuillery (2009, p. 1621).

Figure 4 highlights potential increases in both precision and recall rates from cleaning and parsing techniques, e.g., SSM, Token, 2-gram, and multiple filters (MF) compared to a simple string match (SSM) baseline and with multiple filters. The graph is a variant of the receiver operating characteristic (ROC) curve often used in the data science literature to illustrate the efficiency of binary classifiers as attributes in the model are varied. The area under the curve (AUC) statistic, calculated from the ROC curve, to compare model fit to the data as attributes are varied (Efron & Hastie, 2016, p. 386). Preferred models have AUC values near one; values around 0.5 are no better than chance. In Figure 4 the characteristic curve of an efficient classifier would hug the top right corner of the graph (James et al., 2013). Taken together, this suggests that scholars have good reason to insist on the importance of the clean and parse stage (Magerman et al., 2006) and the SSM should only be used as a basic heuristic benchmark to evaluate precision rate and recall rate gains from more sophisticated techniques. These results should encourage creativity in the use of MF because their availability and influence may differ across data sources using the techniques outlined above and elaborated further below.

[9] The foregoing is sufficient to interpret the graph but more details about its construction are as follows: The four possible results from a binary classifier are: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Recall rate is defined as: (TP+TN) /(TP+TN+FN), where TP and TN are true positive and true negative results (correct recall), respectively, and FN is false negative (Type I error). Precision rate is defined as: (TP+TN)/(TP+TN+FP), where FP is false positive (Type II error).

[10] The points labeled A to D in the original are peripheral to the argument here but can be interpreted as follows: A= 95% precision with a 5% increase in recall compared to SSM+MF; B = 82% precision with an 11% increase in recall compared to SSM; C = 94% precision with a 12% increase in recall compared to SSM+MF and D = 93% precision with a 13% increase in recall compared to SSM+MF.

Theory - The Three Stages

There are considerable methodological problems with creating a nascent person identifier for bibliographic data from different sources (Raffo & Lhuillery, 2009). *First*, a consensus is still emerging on the 'best' procedures to disambiguate name data from two or more sources. *Second*, all name disambiguation procedures involve some *ad hoc* choices that influence the final results. *Third*, while full disclosure of *ad hoc* choices is commendable, what is needed is some concept of reliability and validity for nascent person identifiers. While there is an emerging consensus that different disambiguation techniques can be compared using precision–recall points and curves similar to those in Figure 4, this is not a silver bullet for the methodological problems. In this section, we review and integrate some of the ingenious disambiguation techniques developed in the literature towards solving the nascent person identifier puzzle. To explain where consensus is emerging, we extend Pezzoni and colleagues (2014) with a series of examples around five cartoon characters using a selection of attributes that appear in the literature. The cartoon characters not only make our theoretical points more concrete but also avoid getting overwhelmed with actual data, while discussing theory. We provide examples using our data in later sections. Figure 5 provides canonical attribute data, i.e., consistent and complete, for each of the five characters that we will compare and contrast to the often incomplete, and sometimes inconsistent, data from bibliographic LOD sources.

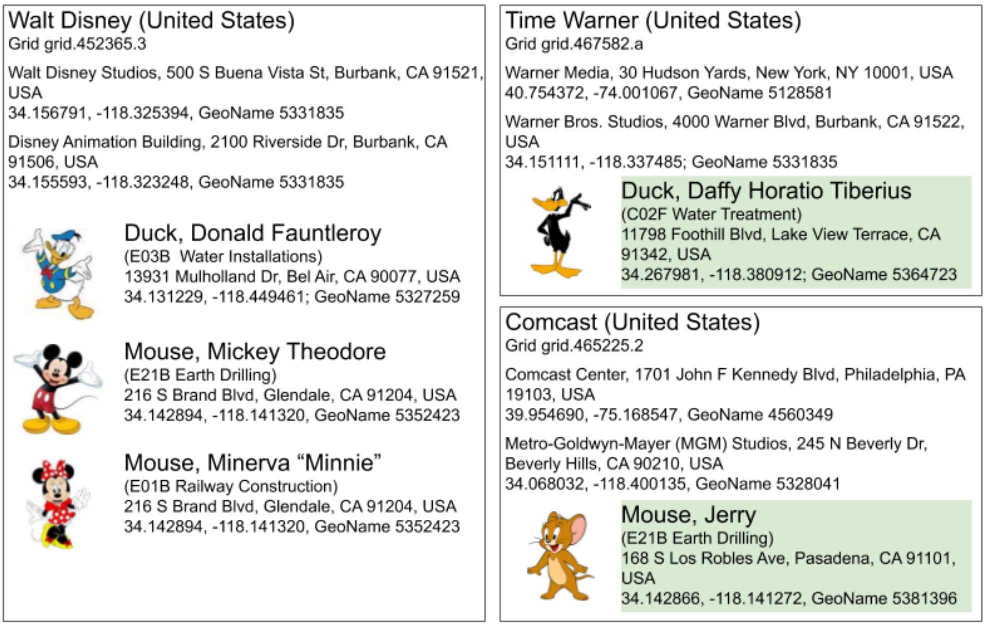
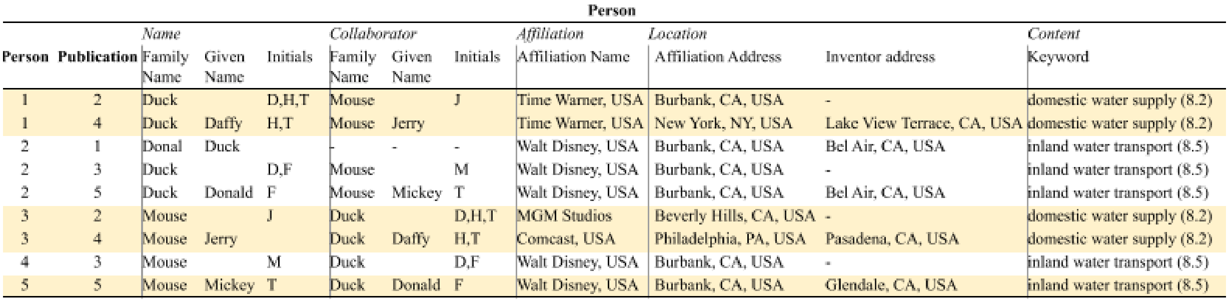

Walt Disney (United States)
Grid grid.452365.3
Walt Disney Studios, 500 S Buena Vista St, Burbank, CA 91521, USA
34.156791, -118.325394, GeoName 5331835
Disney Animation Building, 2100 Riverside Dr, Burbank, CA 91506, USA
34.155593, -118.323248, GeoName 5331835
Duck, Donald Fauntleroy
(E03B Water Installations)
13931 Mulholland Dr, Bel Air, CA 90077, USA
34.131229, -118.449461; GeoName 5327259
Mouse, Mickey Theodore
(E21B Earth Drilling)
216 S Brand Blvd, Glendale, CA 91204, USA
34.142894, -118.141320, GeoName 5352423
Mouse, Minerva "Minnie"
(E01B Railway Construction)
216 S Brand Blvd, Glendale, CA 91204, USA
34.142894, -118.141320, GeoName 5352423
Time Warner (United States)
Grid grid.467582.a
Warner Media, 30 Hudson Yards, New York, NY 10001, USA
40.754372, -74.001067, GeoName 5128581
Warner Bros. Studios, 4000 Warner Blvd, Burbank, CA 91522, USA
34.151111, -118.337485; GeoName 5331835
Duck, Daffy Horatio Tiberius
(C02F Water Treatment)
11798 Foothill Blvd, Lake View Terrace, CA 91342, USA
34.267981, -118.380912; GeoName 5364723
Comcast (United States)
Grid grid.465225.2
Comcast Center, 1701 John F Kennedy Blvd, Philadelphia, PA 19103, USA
39.954690, -75.168547, GeoName 4560349
Metro-Goldwyn-Mayer (MGM) Studios, 245 N Beverly Dr, Beverly Hills, CA 90210, USA
34.068032, -118.400135, GeoName 5328041
Mouse, Jerry
(E21B Earth Drilling)
168 S Los Robles Ave, Pasadena, CA 91101, USA
34.142866, -118.141272, GeoName 5381396

Figure 5. Canonical attributes for *nascent* person identifier examples.

The Figure shows canonical names for five people, i.e., inventors, authors, and other attributes, i.e., collaboration, affiliation, location, and content. Based on Pezzoni and colleagues (2014, p. 483).

Clean and Parse

The objective of the clean and parse stage is to reduce 'noise' in the source text strings uploaded to attribute fields of preliminary tables, while preserving residual text fragments that may be useful at the filter stage. For example, in Figure 6 for person #2 and publication #5: the person name text string Duck, Donald F., would be parsed into 'Duck' and 'Donald' for the family name and given name fields, respectively. However, the middle name or initial, 'Fauntleroy' or 'F.', should also be retained in a separate field because they can be used at the filtering stage to help disentangle homonyms, i.e, different people with the same or very similar family and given names. Name text strings may also contain a surname prefix, e.g., De, Mc, Mac, Van, which can be rejoined to the family name string using a 'prefix' lookup table (not shown). In addition, if there is a prefix or suffix in the name field, e.g., Prof. or Dr.; Ph.D., Jnr., or III, respectively, then these can also be held in separate fields (also not shown), and introduced later as filters, similar to initials. Figure 6 illustrates the fundamental challenge of name disambiguation, i.e., to use the available attributes to classify publications #1 to 5 to the correct *nascent* identifier, person #1 to 5. In the examples that follow, we retain illustrative attributes that generalise to important additional attribute types, e.g., collaborator, affiliation, location, and content.



Person

		Name			*Collaborator*			*Affiliation*	*Location*		*Content*
Person	**Publication**	Family Name	Given Name	Initials	Family Name	Given Name	Initials	Affiliation Name	Affiliation Address	Inventor address	Keyword
1	2	Duck		D,H,T	Mouse		J	Time Warner, USA	Burbank, CA, USA	-	domestic water supply (8.2)
1	4	Duck	Daffy	H,T	Mouse	Jerry		Time Warner, USA	New York, NY, USA	Lake View Terrace, CA, USA	domestic water supply (8.2)
2	1	Donal	Duck		-	-	-	Walt Disney, USA	Burbank, CA, USA	Bel Air, CA, USA	inland water transport (8.5)
2	3	Duck		D,F	Mouse		M	Walt Disney, USA	Burbank, CA, USA	-	inland water transport (8.5)
2	5	Duck	Donald	F	Mouse	Mickey	T	Walt Disney, USA	Burbank, CA, USA	Bel Air, CA, USA	inland water transport (8.5)
3	2	Mouse		J	Duck		D,H,T	MGM Studios	Beverly Hills, CA, USA	-	domestic water supply (8.2)
3	4	Mouse	Jerry		Duck	Daffy	H,T	Comcast, USA	Philadelphia, PA, USA	Pasadena, CA, USA	domestic water supply (8.2)
4	3	Mouse		M	Duck		D,F	Walt Disney, USA	Burbank, CA, USA	-	inland water transport (8.5)
5	5	Mouse	Mickey	T	Duck	Donald	F	Walt Disney, USA	Burbank, CA, USA	Glendale, CA, USA	inland water transport (8.5)

Figure 6. Person table of person-publication records, emphasising ideal *result*.

The Figure shows the *canonical* version of the person table. Person table contains person-publication records with a nascent person identifier and attribute fields to help disambiguate people despite several fields having missing data. The person table is assembled from preliminary tables, e.g., collaborator, affiliation, location, and content. The person table is the input for the match and filter stages that follow. Based on Li and colleagues (2014, p. 944).

Figure 6 shows the person table where the input data for the disambiguation process is assembled. The Figure also shows that the unit of analysis in our disambiguation problem is a *person-publication* record that corresponds to one row of the person table and is indexed by the nascent person identifier and publication identifier keys. Each combination of *person* and *publication* occurs only once. However, both *person* and *publication* may appear in several different records. Obviously, a nascent person identifier appears every time the person collaborates on a publication, i.e., Figure 6 shows ID #2 appears three times: for publications #1, 3, and 5. Perhaps less obviously, a publication identifier appears for every collaborator involved on a publication, e.g., Figure 6 shows publication #4 appears twice: for person #1 and 3. Consequently, the EPO LOD contributes **9.9 million** inventor-patent records to the IP LodB database from the 1830s to the present.[11,12] The SN LOD contributes another **34.9 million** author-article (book, book chapter) records, for the same period.[13,14] Thus, the person

[11] The EPO LOD contains data for 9,899,524 inventor-patent records.

[12] In addition, the EPO LOD contains data for 4,654,249 applicant-patent records and 4,283,945 agent-patent records.

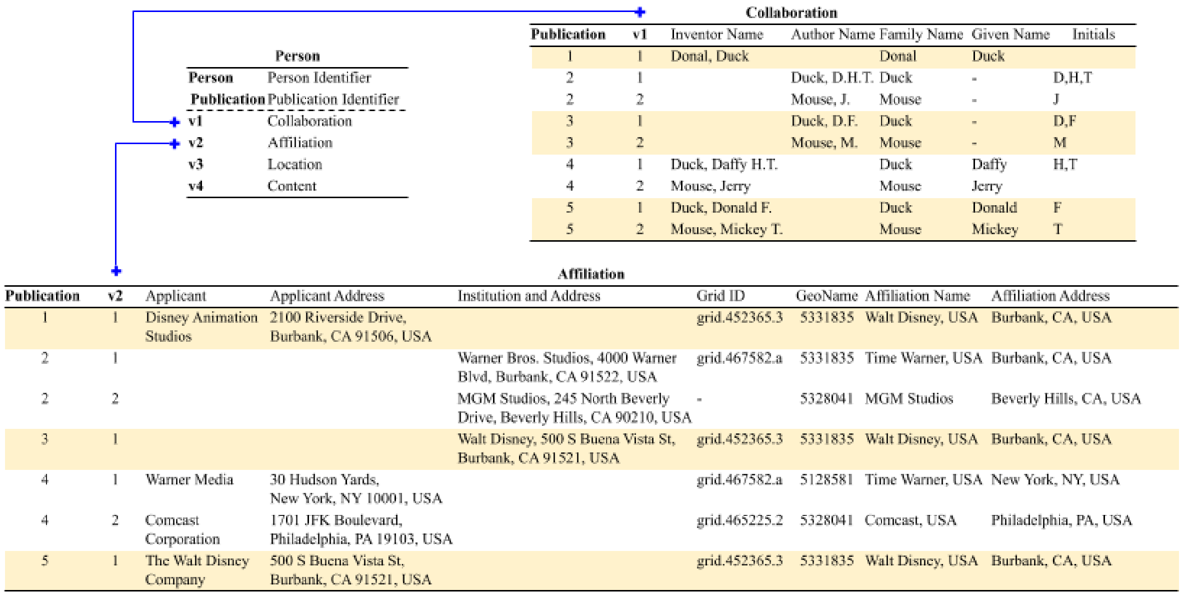
[13] The SN LOD contains data for 34,889,468 author-article (book, book chapter) records, which include 20,855,322 author-article records, 188,621 author-book records, and 13,845,525 author-book chapter records.

[14] In addition, the SN LOD contains data for 487,810 author-grant recipients.

table in the IP LodB database contains **44.8 million** person-publication records because it merges data from both sources. The attributes in the person table shown in Figure 6 are labeled following the process outlined in Figure 2 and the parsing and cleaning techniques in preliminary tables in Figures 7 and 10, below.

Name and Collaborator

Each person-publication record contains text string indicators for person attributes, including their name and the names for their collaborators, affiliations, etc. Obviously, the name field is the *most* clearly disambiguating attribute for any person. The collaboration table in Figure 7 illustrates that the raw text strings from inventor name and author name fields in the source data can be decomposed into family name and given name sub-strings, also called tokens, to help reveal a person's canonical name. Typically, in bibliographic data, family name is recorded first and given name is recorded after a comma, e.g., "Duck, Donald", so a good initial working assumption is that the first token contains the family name (or at least part of it, in the case of double or triple family names) and the part after the 'comma' contains the given name (or part of it, in the case of second names or initials). The collaboration table in Figure 7 illustrates that records from SN LOD are systematically missing given names because author name fields in the source data only contain initials after the 'comma'. Low resolution author name fields, i.e., given names represented by initials only, are a major disadvantage for disambiguating journal articles (book, book chapters), compared to patent documents that typically have high resolution given names (Li et al., 2014).



Person

Person	Person Identifier
Publication	Publication Identifier
v1	Collaboration
v2	Affiliation
v3	Location
v4	Content

Collaboration

Publication	v1	Inventor Name	Author Name	Family Name	Given Name	Initials
1	1	Donal, Duck		Donal	Duck	
2	1		Duck, D.H.T.	Duck	-	D,H,T
2	2		Mouse, J.	Mouse	-	J
3	1		Duck, D.F.	Duck	-	D,F
3	2		Mouse, M.	Mouse	-	M
4	1	Duck, Daffy H.T.		Duck	Daffy	H,T
4	2	Mouse, Jerry		Mouse	Jerry	
5	1	Duck, Donald F.		Duck	Donald	F
5	2	Mouse, Mickey T.		Mouse	Mickey	T

Affiliation

Publication	v2	Applicant	Applicant Address	Institution and Address	Grid ID	GeoName	Affiliation Name	Affiliation Address
1	1	Disney Animation Studios	2100 Riverside Drive, Burbank, CA 91506, USA		grid.452365.3	5331835	Walt Disney, USA	Burbank, CA, USA
2	1			Warner Bros. Studios, 4000 Warner Blvd, Burbank, CA 91522, USA	grid.467582.a	5331835	Time Warner, USA	Burbank, CA, USA
2	2			MGM Studios, 245 North Beverly Drive, Beverly Hills, CA 90210, USA	-	5328041	MGM Studios	Beverly Hills, CA, USA
3	1			Walt Disney, 500 S Buena Vista St, Burbank, CA 91521, USA	grid.452365.3	5331835	Walt Disney, USA	Burbank, CA, USA
4	1	Warner Media	30 Hudson Yards, New York, NY 10001, USA		grid.467582.a	5128581	Time Warner, USA	New York, NY, USA
4	2	Comcast Corporation	1701 JFK Boulevard, Philadelphia, PA 19103, USA		grid.465225.2	5328041	Comcast, USA	Philadelphia, PA, USA
5	1	The Walt Disney Company	500 S Buena Vista St, Burbank, CA 91521, USA		grid.452365.3	5331835	Walt Disney, USA	Burbank, CA, USA

Figure 7. Preliminary tables for collaboration and affiliation.

The Figure shows the collaboration and affiliation tables are preliminary tables, to the person table, where the source bibliographic data are cleaned and parsed. The collaboration table parses name text strings into family name and given name, and other tokens, which are carried forward to the person name and collaborator name fields in the person table. The affiliation table parses both applicant and institution name and address strings into tokens to validate with GRID and GeoNames identifiers, which are carried forward to the affiliation name and affiliation address fields in the person table, respectively. The GeoNames come bundled with the GRID identifiers. Based on Li and colleagues (2014, p. 944).

The collaboration table in Figure 7 illustrates that name text strings also indicate collaborators for a person. Perhaps less obviously, while no additional attribute is enough on its own, a person's network of collaborators is their *second* most clearly disambiguating attribute, after their name. Thus, if two people have the same or very similar names and one or more collaborators with the same or very similar names, then the two person-publication records *probably* refer to the same person. The distinction between person name and collaborator name is made explicit when cleaned and parsed text strings are carried forward from the preliminary collaboration table, in Figure 7, to the person table, in Figure 6. Routine cleaning techniques include the removal of capital letters, punctuation, non-alpha-numeric symbols, accented characters, and double spaces. Figures 7 and 10 provide a schematic of the data preparation process where preliminary tables are created from source data and then, after data is cleaned and parsed, they are brought forward to the person table for match and filter operations.

Decomposing text strings into tokens also allows person-publication records to be reorganized into different 'groups' using lexical distance, e.g., Levenstein distance (LD), 2-gram distance (2G), which facilitate alternative match and filter operations. The tokens extracted from simple text strings are sorted alphabetically, without distinguishing between family names and given names.[15] The 2G lexical distance between two text strings of different lengths, normalized by the total length of the strings, can then be evaluated. Equation 1 can compute the distance between consecutive tokens on the list, i.e., difference between row n and row n+1:

$$2G\left(t_1, t_2\right) = \frac{\sqrt{\sum_{i=aa}^{zz(N)} \left(G_{1i} - G_{2i}\right)^2}}{num\left(t_1\right) + num\left(t_2\right)} \tag{1}$$

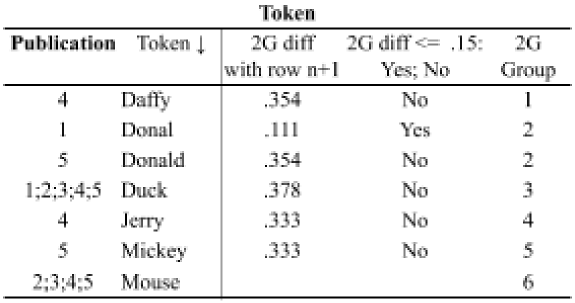
where $G_{1\psi\psi}$ and $G_{2\psi}$ are the number of occurrences the *i*th bigram appears in tokens t_1 and t_2, respectively, *num*(t_1) and *num*(t_2) are the number of characters in tokens t_1 and t_2, respectively, and N is the number of possible combinations of two consecutive letters, i.e., bigrams, in the alphabet of choice, e.g., $N \times \psi = 26 \times 25 = 650$, for the plain ASCII character set. For example, take the tokens "Donal" as t_1 and "Donald" as t_2 from Figure 8. The bigram sets for t_1 and t_2 are: (Do,on,na,al) and (Do,on,na,al,ld), respectively. Applying equation 1 returns:

$$\begin{aligned} 2G\left(t_1, t_2\right) &= \frac{\sqrt{(1-1)_{Do}^2 + (1-1)_{on}^2 + (1-1)_{na}^2 + (1-1)_{al}^2 + (0-1)_{ld}^2}}{4+5} \\ &= \frac{1}{9} = .111 \end{aligned}$$

Once all 2G(t_1,t_2) distances are computed, consecutive tokens can be assigned to 'groups' using some threshold value for 2G lexical distance, e.g., $\delta = 0.15$, as follows:

[15] Yields about 500k tokens, excluding one and two character tokens (Pezzoni et al., 2014, p. 481).

1) From the top of the list of tokens, sorted alphabetically, the first token is assigned to group 1.
2) Then, if the 2G lexical distance between the token in row 1 and the token in row 2 is less than the threshold, e.g., 0.15, then the token in row 2 is also in group 1; otherwise a new group is created, i.e., group 2.
3) The same procedure is applied to tokens in row n and n+1 to the end of the list.

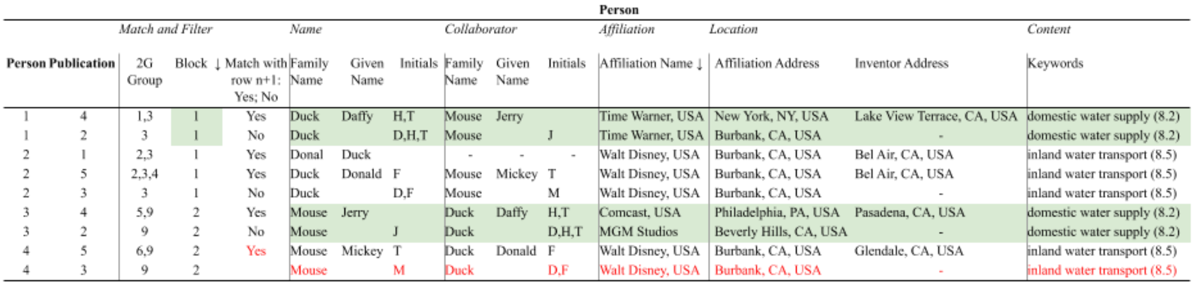


Token				
Publication	Token ↓	2G diff with row n+1	2G diff <= .15: Yes; No	2G Group
4	Daffy	.354	No	1
1	Donal	.111	Yes	2
5	Donald	.354	No	2
1;2;3;4;5	Duck	.378	No	3
4	Jerry	.333	No	4
5	Mickey	.333	No	5
2;3;4;5	Mouse			6

Figure 8. Tokens and lexical distance to manage misspellings and inversions in family and given names.

The figure shows how tokens help manage minor misspellings and inversions in family names and given names using the decomposition text string into tokens of three or more characters. The tokens are sorted alphabetically, irrespective of whether they are family names and given names. The lexical distance from one token to the next is computed, e.g., 2G = 2-gram distances, to determine whether the tokens are in the same group. Based on Pezzoni and colleagues (2014, p. 483).

Figure 8 illustrates the results of token decomposition, lexical distance, and group assignment, for the family name and given name text strings in the collaboration table from Figure 7. The groups are the same in this example but that doesn't hold in general. Figure 9 is an alternative representation of the person table in Figure 6, and illustrates how the person-publication records have been sorted using 2G groups to form blocks 1 and 2. These blocks are considerably more permissive than the blocks allowed by SSM, where *none* of these nine person-publication records would have been considered as potential matches. In addition, the Figure presents a coding structure for attributes in a more amenable format for use with a binary classifier. These token, lexical distance, and coding techniques represent major advances in name disambiguation and we return to how they are used in the match and filter stages below.



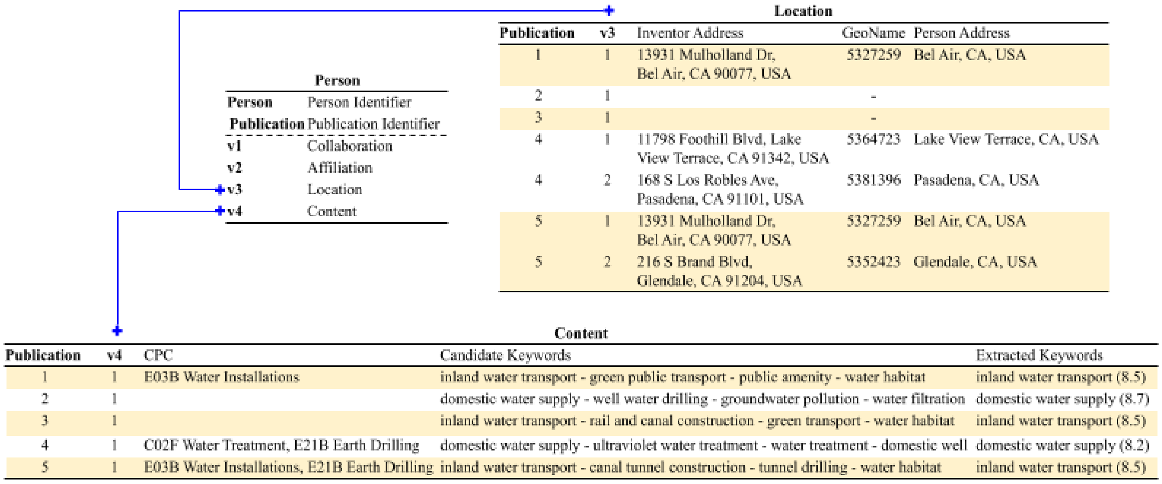
Person														
		Match and Filter			*Name*			*Collaborator*			*Affiliation*	*Location*		*Content*
Person	Publication	2G Group	Block ↓	Match with row n+1: Yes; No	Family Name	Given Name	Initials	Family Name	Given Name	Initials	Affiliation Name ↓	Affiliation Address	Inventor Address	Keywords
1	4	1,3	1	Yes	Duck	Daffy	H,T	Mouse	Jerry		Time Warner, USA	New York, NY, USA	Lake View Terrace, CA, USA	domestic water supply (8.2)
1	2	3	1	No	Duck		D,H,T	Mouse		J	Time Warner, USA	Burbank, CA, USA	-	domestic water supply (8.2)
2	1	2,3	1	Yes	Donal	Duck		-	-	-	Walt Disney, USA	Burbank, CA, USA	Bel Air, CA, USA	inland water transport (8.5)
2	5	2,3,4	1	Yes	Duck	Donald	F	Mouse	Mickey	T	Walt Disney, USA	Burbank, CA, USA	Bel Air, CA, USA	inland water transport (8.5)
2	3	3	1	No	Duck		D,F	Mouse		M	Walt Disney, USA	Burbank, CA, USA	-	inland water transport (8.5)
3	4	5,9	2	Yes	Mouse	Jerry		Duck	Daffy	H,T	Comcast, USA	Philadelphia, PA, USA	Pasadena, CA, USA	domestic water supply (8.2)
3	2	9	2	No	Mouse		J	Duck		D,H,T	MGM Studios	Beverly Hills, CA, USA	-	domestic water supply (8.2)
4	5	6,9	2	Yes	Mouse	Mickey	T	Duck	Donald	F	Walt Disney, USA	Burbank, CA, USA	Glendale, CA, USA	inland water transport (8.5)
4	3	9	2		Mouse		M	Duck		D,F	Walt Disney, USA	Burbank, CA, USA	-	inland water transport (8.5)

Figure 9. Person table of person-publication records, emphasising heuristic *process.*

The Figure shows a *process* version of the person table. Person-publication records with the same or similar name text strings and one or more attributes that are also the same or similar probably refer to the same person. Sorting the data using, for example, *first*, 2G groups, i.e., block 1 and 2, and second, affiliation name, helps to see patterns in the data. Such sorting of the data is directly analogous to the filter stage of the name disambiguation process, after initial record matches have been accepted. Note that all the records shown in the Figure would be *split*, i.e., false negatives, using the SSM decision rule. Also, note that the last record is lumped, i.e., a false positive, with the second last record because the last record should refer to Minnie Mouse. Based on Li and colleagues (2014, p. 944).

Affiliation and Location

Affiliation names also suffer from misspelling and name form variation, e.g., Warner Bros. Studios vs Time-Warner (United States), especially for a subsidiary with a completely different name, i.e., that has no words in common, e.g., MGM Studios vs Comcast (United States). The source data for applicant and institution name and address fields shown in the affiliation table in Figure 7 can also be decomposed into tokens to help reveal affiliation name and location. However, the disambiguation of affiliation is compounded in SN LOD because the institution name and institution address text strings are in the same field, shown in the affiliation table in Figure 7. However, an alternative strategy is to validate affiliation name tokens using the GRID database of research organizations or HAN database of patent assignees to disambiguate institutions for authors and applicants for inventors, respectively. The affiliation problem for the SN LOD source data is ameliorated somewhat because they provide links (URIs) to records in GRID, which usually include GeoNames, and cover 86% of author-article records. Nevertheless, while the token ‘Warner’ resolves to its parent Time-Warner (United States) in the GRID database, the token ‘MGM’ does not resolve to Comcast (United States), so the strategy is not perfect. Further cleaning and parsing procedures for ambiguous affiliation name and affiliation address fields are described further below.



Location

Publication	v3	Inventor Address	GeoName	Person Address
1	1	13931 Mulholland Dr, Bel Air, CA 90077, USA	5327259	Bel Air, CA, USA
2	1		-	
3	1		-	
4	1	11798 Foothill Blvd, Lake View Terrace, CA 91342, USA	5364723	Lake View Terrace, CA, USA
4	2	168 S Los Robles Ave, Pasadena, CA 91101, USA	5381396	Pasadena, CA, USA
5	1	13931 Mulholland Dr, Bel Air, CA 90077, USA	5327259	Bel Air, CA, USA
5	2	216 S Brand Blvd, Glendale, CA 91204, USA	5352423	Glendale, CA, USA

Person

Person	Person Identifier
Publication	Publication Identifier
v1	Collaboration
v2	Affiliation
v3	Location
v4	Content

Content

Publication	v4	CPC	Candidate Keywords	Extracted Keywords
1	1	E03B Water Installations	inland water transport - green public transport - public amenity - water habitat	inland water transport (8.5)
2	1		domestic water supply - well water drilling - groundwater pollution - water filtration	domestic water supply (8.7)
3	1		inland water transport - rail and canal construction - green transport - water habitat	inland water transport (8.5)
4	1	C02F Water Treatment, E21B Earth Drilling	domestic water supply - ultraviolet water treatment - water treatment - domestic well	domestic water supply (8.2)
5	1	E03B Water Installations, E21B Earth Drilling	inland water transport - canal tunnel construction - tunnel drilling - water habitat	inland water transport (8.5)

Figure 10. Preliminary tables for location and content

The Figure shows the location and content tables are preliminary tables, to the person table, where the source bibliographic data are cleaned and parsed. The location table parses inventor address text strings into city, state, and country, address tokens to validate with GeoNames. The content table parses titles and abstracts from both patent documents and journal articles keywords using the rapid automatic keyword extraction (RAKE) tool that are carried forward to the keyword field in the person table. Based on Li and colleagues (2014, p. 944).

High resolution address strings cleaned and parsed into street address tokens, i.e., street number, street name, postcode, city, state, country, indicate “location” using map software, e.g., Google Maps, because street address tokens can be resolved to exact ‘rooftop’ coordinates, e.g., 34.131229, -118.449461 decimal degrees (DD) for the latitude and longitude for Donald’s hilltop mansion overlooking Stone Canyon Reservoir in Bel Air, shown by a blue pin in Figure 11. Thus, if two inventors have the same or similar names and the same or similar addresses, then the two records *very probably* refer to the same person (Morrison et al., 2017). For example, in the collaboration table in Figure 7 publication #1 by Donal, Duck, and publication #5 by Duck, Donald F., have exactly the same address in the location table in Figure 10, so despite differences in name text strings the two records are likely to refer to the same person. In addition, coordinates are robust to differences in address tokens (Li et al., 2014), i.e., missing street number or misspelling in street name tokens, e.g., without the street number on Mulholland Drive the street address tokens resolve to a location only 300m away from the exact address.[16] Thus, and perhaps more importantly, coordinates are also indicators of “closeness” between locations (Li et al., 2014), i.e., inventors and their collaborators and inventors and applicants, e.g., Figure 11 illustrates Donald’s 21 km commute to the Disney Animation Building in Burbank (16 to 22 mins at rush hour), which is a little further away than his main collaborators, Mickey and Minnie, but only just above the mean commute for metropolitan areas in the United States.

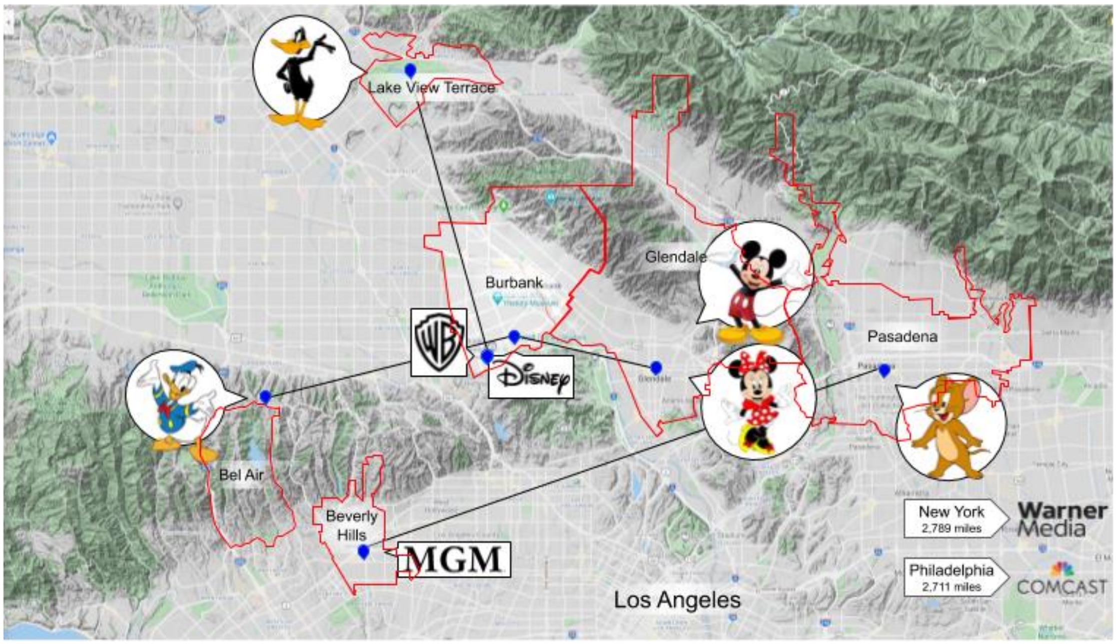


Figure 11. Concepts of “location” and “closeness” for both inventor and affiliation

The Figure shows inventor and affiliation addresses in the Los Angeles area. Two possible inferences: 1) If two people have the same or similar names and their affiliation has the same or similar names or the same or similar address, then the two records *probably* refer to the same person. 2) If two people have the same or similar names and the inventor and affiliation have addresses within about 20km, then the two records also *probably* refer to the same person. Based on Morrison and colleagues (2017, p. 4). Map data adapted from Google Maps.

Street number and street name are typically not available in the United States from USPTO (Morrison et al., 2017), although they are typically available for inventors in Europe

[16] 34.131747, -118.451992 decimal degrees (DD)

from PATSTAT but not EPO LOD. If street number or street name tokens are not available, then postcode address tokens, i.e., postcode, city, state, country, resolve to an attractive second best location because postcode divides large cities into dozens of smaller segments. However, with other attributes, e.g., collaborator, affiliation, content, even city address tokens, i.e., city, state, country, resolve to an adequate third best location measure (Li et al., 2014). Figure 11 illustrates that GeoNames resolve city address tokens to areas marked by red lines, e.g., city address tokens, “Burbank, CA, USA”, to 5331835. Obviously, any city address token contains many blue pins, e.g., both Disney and Warner Bros. have a campus in Burbank. Unfortunately, GeoNames cannot resolve postcode address tokens, i.e., CA 91521, Burbank, USA, to postcode granularity because Disney and Warner Bros. have different postcodes, CA 91521 and CA 91522, respectively, even though they are only 2.2 km apart. So, postcode address tokens are considerably more accurate than city address tokens but GeoNames cannot resolve them. Google Maps can resolve postcode address tokens for a fee but not with the resources available to this project.

More importantly, author address details are generally not available for journal articles. Consequently, institution name and address become the only indicators of either “location” or “closeness” for records from SN LOD. Thus, institution name and address become relatively more important attributes in our project than previous work on nascent person identifiers would suggest. Similar to other publishers, the SN LOD typically provides institution name and address for at least the first author of journal articles. Thus, if two people have the same or similar names and their affiliation has the same or similar name or the same or similar address, then the two records *probably* refer to the same person, e.g., in the affiliation table in Figure 7, publication #1 Disney Animation Studio is the applicant attribute and publication #3 Walt Disney is the institution but their addresses are only 400m appart, so despite differences in text strings for both affiliation name and address the two records are very likely to refer to the same organization.

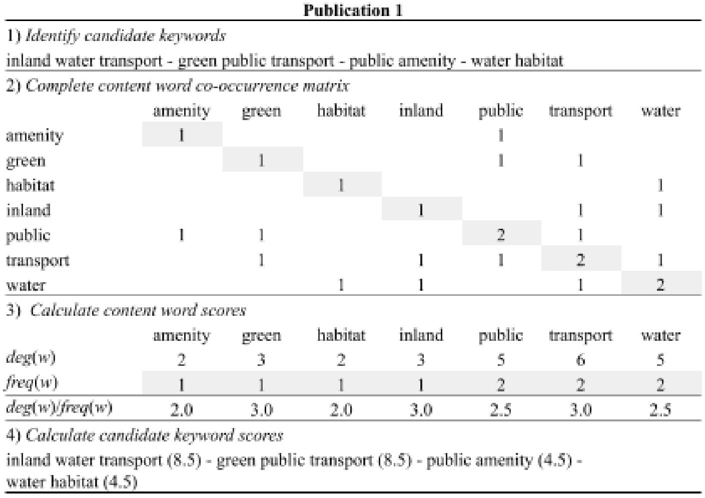
Nevertheless, some records can have institution addresses close by inventor address but far away from applicant address, so the organization affiliation table does not supersede the location table. For example, in the affiliation table in Figure 6, publication #4 has Comcast Corporation with an address in Philadelphia as one applicant, which is 2,711 miles away from either of the institutions for publication #2, i.e., Warner Bros. Studios with an address in Burbank or MGM Studios with an address in Beverly Hills. However, the inventor, Jerry Mouse, has an address in Pasadena, which is within commuting distance for either of the latter institutions. Thus, if two people have the same or similar names and the affiliation and inventor have addresses within about 20km, then the two records *also probably* refer to the same person. Thus, affiliation and location become the *third* and *forth* most clearly disambiguating attributes for people in our data, i.e., after name and collaborator.

Content

Keywords are a sequence of one or more words that provide a compact indication of a document’s content (Rose et al., 2010). For patent documents, the content table in Figure 10 contains examples from an extensive taxonomy of technology classes, i.e., keywords indicating inventor expertise, labeled cooperative patent classification (CPC). If two inventors have the same or similar names and one or more of the same or similar CPC fields, then the two records *probably* refer to the same person. However, no equivalent keyword taxonomy exists for journal articles. So, to provide indicators of content across both patent documents and journal articles, we identify keywords from individual publication titles and abstracts using the rapid automatic keyword extraction (RAKE) procedure. RAKE is an unsupervised, domain and language independent procedure designed to identify keywords from titles and abstracts of individual publications (Rose et al., 2010). Keywords often contain multiple content words but rarely contain punctuation, e.g., commas, semi-colons, or stop words, e.g., *the*, *and*, *of*, and other words with minimal content. Building on this insight, RAKE uses three input parameters; word delimiters, phrase delimiters, and stop words, to parse text strings into candidate keywords. Candidate keywords are contiguous sequences of content words that appear at least twice in the title or abstract text in the same publication and in the same order (Rose et al., 2010). Thus, RAKE can adapt to the style and content of individual publications to provide a fine-grained and nuanced list of candidate keywords.

We apply RAKE to both EPO and SN LOD sources, but only one block at a time, i.e., for one family name and given name initial, e.g., Duck, D., and, thus, parse a list of candidate keywords, as shown in the content table of Figure 10. The longish list of *candidate keywords* for individual publications is whittled down to a much shorter list of *extracted keywords* for the collection of documents in a block, also shown in the content table of Figure 10. The keywords may indicate very different contents, e.g., railway construction and water treatment, or deceptively similar domains, e.g., water treatment and water installations. However, if RAKE is specified correctly, extracted keywords can allow us to discriminate between technical domains. Thus, if two people have the same or similar names and one or more of the same or similar extracted keywords, then the two records *probably* refer to the same person. Table 1 specifies the first part of the RAKE process where an initial list of candidate keywords are parsed from the title and abstract to indicate the content of individual publications. Table 2 specifies the second part of the RAKE process where the initial list is whittled down to a shorter list of extracted keywords that indicate the contents for a heterogeneous collection.

Table 1. Keywords for individual publications based on RAKE



Publication 1							
1) *Identify candidate keywords*							
inland water transport - green public transport - public amenity - water habitat							
2) *Complete content word co-occurrence matrix*							
	amenity	green	habitat	inland	public	transport	water
amenity	1				1		
green		1			1	1	
habitat			1				1
inland				1		1	1
public	1	1			2	1	
transport		1		1	1	2	1
water			1	1		1	2
3) *Calculate content word scores*							
	amenity	green	habitat	inland	public	transport	water
deg(*w*)	2	3	2	3	5	6	5
freq(*w*)	1	1	1	1	2	2	2
deg(*w*)/*freq*(*w*)	2.0	3.0	2.0	3.0	2.5	3.0	2.5
4) *Calculate candidate keyword scores*							
inland water transport (8.5) - green public transport (8.5) - public amenity (4.5) - water habitat (4.5)							

The Table shows the RAKE procedure for individual publications, e.g., publication 1, in four stages: 1) identify candidate keywords parsed from title and abstract. 2) Complete content word co-occurrence matrix. 3) Calculate content word scores for degree *deg*(*w*), frequency *freq*(*w*), and ratio of degree to frequency *deg*(*w*)/*freq*(*w*). 4) Calculate candidate keyword scores using *deg*(*w*)/*freq*(*w*), e.g., inland water transport (8.5), i.e., 3 + 2.5 + 3 = 8.5. The candidate keywords with scores in the top third are carried forward as extracted keywords. Based on Rose and colleagues (2010, pp. 6–8).

Table 1 specifies *four* stages of the RAKE procedure for the first part of the RAKE process where a list of candidate keywords are parsed from the title and abstract to indicate the content of individual publications:

1) *Identify candidate keywords* from title and abstract text strings in the order that they are parsed from the text using three RAKE parameters; word delimiters, phrase delimiters, and stop words.
2) *Complete content word co-occurrence matrix* from the list of candidate keywords.
3) *Calculate content word scores* using: degree *deg*(*w*), the sum of values in each column; frequency *freq*(*w*), the values on the diagonal, and ratio of degree to frequency *deg*(*w*)/*freq*(*w*).
4) *Calculate candidate keyword* scores using the sum of the *deg*(*w*)/*freq*(*w*) score for each content word, e.g., inland water transport (8.5). The candidate keywords with scores in the top third are carried forward as extracted keywords.

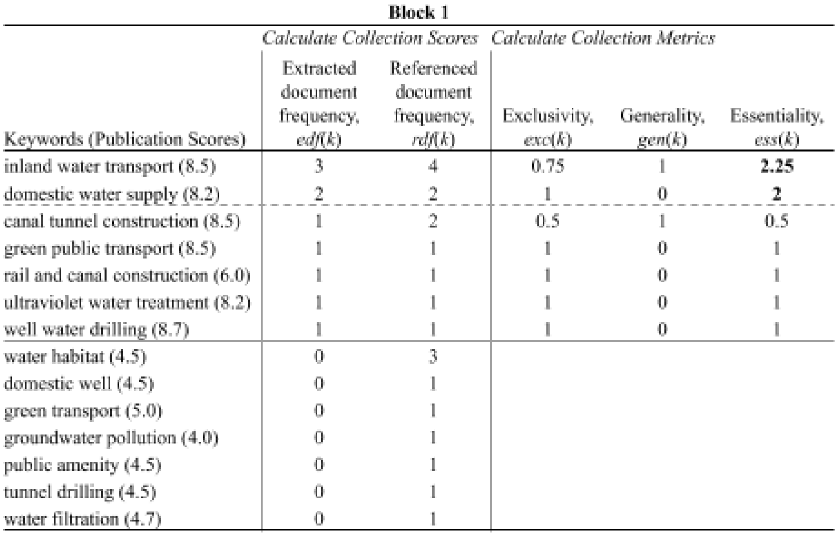
Each content word score has a different emphasis: Degree *deg*(*w*) emphasises words that occur often and in longer candidate keywords, e.g., *deg*(*green*) = 3 is less than *deg*(*transport*) = 6. Frequency *freq*(*w*) emphasises words that occur often regardless of the length of candidate keywords, e.g., *freq*(*green*) = 1 is also less than *freq*(*transport*) = 2. However, the ratio *deg*(*w*)/*freq*(*w*) emphasises words that occur in longer candidate keywords regardless of the frequency of the content word, e.g., *deg*(*green*)/*freq*(*green*) = 3 is equal to *deg*(*transport*)/*freq*(*transport*) =3. Candidate keyword scores are calculated using the sum of the ratio *deg*(*w*)/*freq*(*w*) metric scores for each content word. After all candidate keywords are scored, the top third are carried forward as extracted keywords to indicate the content of individual publications.

For collections of publications, we use the same simple parameters for RAKE with some post estimation arithmetic to evaluate keywords. Candidate keywords should appear in *two* or more publications so keywords can both converge on and discriminate between contents in a collection. The ratio *deg*(*w*)/*freq*(*w*) metric produces fewer and longer candidate keywords with lower frequency across publications, compared to degree *deg*(*w*). The ratio *deg*(*w*)/*freq*(*w*) metric also produces relatively few uninformative keywords, e.g., introduction, method, etc. So, the keywords based on the *deg*(*w*)/*freq*(*w*) ratio have relatively high convergent and discriminant validity. Thus, we use the ratio *deg*(*w*)/*freq*(*w*) metric because it is most likely to produce discriminating keywords indicating individual publication content in heterogeneous collections. For collections, keyword scores are based on values in columns one and two of Table 2:

1) Referenced document frequency *rdf*(*k*) is the number of publications where a keyword was identified as a candidate keyword.
2) Extracted document frequency *edf*(*k*) is the number of publications where a keyword was carried forward as an extracted keyword.

The *rdf*(*k*) and *edf*(*k*) are then used to calculate three metrics for extracted keywords from the collection: exclusivity *exc*(*k*), generality *gen*(*k*), and essentiality *ess*(*k*), the most important of which for our purposes is essentiality *ess*(*k*).

Table 2. Keyword for a collection of publications based RAKE



Block 1

	Calculate Collection Scores		*Calculate Collection Metrics*		
Keywords (Publication Scores)	Extracted document frequency, *edf*(*k*)	Referenced document frequency, *rdf*(*k*)	Exclusivity, *exc*(*k*)	Generality, *gen*(*k*)	Essentiality, *ess*(*k*)
inland water transport (8.5)	3	4	0.75	1	**2.25**
domestic water supply (8.2)	2	2	1	0	**2**
canal tunnel construction (8.5)	1	2	0.5	1	0.5
green public transport (8.5)	1	1	1	0	1
rail and canal construction (6.0)	1	1	1	0	1
ultraviolet water treatment (8.2)	1	1	1	0	1
well water drilling (8.7)	1	1	1	0	1
water habitat (4.5)	0	3			
domestic well (4.5)	0	1			
green transport (5.0)	0	1			
groundwater pollution (4.0)	0	1			
public amenity (4.5)	0	1			
tunnel drilling (4.5)	0	1			
water filtration (4.7)	0	1			

The table shows the RAKE procedure for a collection of publications, e.g., block 1, in two stages: 1) Calculate collection scores for the number of publications where candidate keywords were a) referenced *rdf*(*k*) and b) extracted *edf*(*k*). 2) Calculate collection metrics to evaluate extracted key words: exclusivity *exc*(*k*) = *edf*(*k*)/*rdf*(*k*), essentiality *ess*(*k*) = *exc*(*k*)×*edf*(*k*), and generality *gen*(*k*) = *rdf*(*k*)×(1-*exc*(*k*)). Based on Rose and colleagues (2010, pp. 16–18).

Table 2 shows the arithmetic to evaluate candidate and extracted keywords shown in the content table of Figure 10 and the extracted keywords brought forward to the person table in Figure 6 and 9. Three collection metrics in columns three, four, and five of Table 2 are calculated and interpreted as follows:

1) Exclusive *exc*(*k*) keywords are those that are extracted from all publications where they are identified as candidate keywords, e.g., *exc*(*domestic water supply*) = 1 and is more exclusive than *exc*(*inland water transport*) = 0.75. Exclusivity *exc*(*k*) is defined as:

$$exc\,(k) = \frac{edf\,(k)}{rdf\,(k)} \tag{2}$$

   where *edf*(*k*) is the number of publications where the keyword was extracted and *rdf*(*k*) is the number of publications where the keyword was identified as a candidate.

2) General *gen*(*k*) keywords, in contrast, are those that are extracted from relatively few publications where they are identified as candidate keywords, e.g., *exc*(*inland water transport*) = 1 and is more general than *exc*(*domestic water supply*) = 0. Generality *gen*(*k*) is defined as:

$$gen\,(k) = rdf\,(k) \times \left(1 - \frac{edf\,(k)}{rdf\,(k)}\right) \tag{3}$$

   where *edf*(*k*) and *rdf*(*k*) are the same as above.

3) Essential *ess*(*k*) keywords are those that are not only relatively exclusive *exc*(*k*) but also extracted frequently *edf*(*k*), indicating important contents in a collection of publications, e.g., *ess*(*inland water transport*) = 2.25 and is more essential than *exc*(*domestic water supply*) = 2. Essentiality *ess*(*k*) is defined as:

$$ess\,(k) = exc\,(k) \times edf\,(k) \tag{4}$$

   where *exc*(*k*) and *edf*(*k*) are the same as above.

Match

Put simply, the objective of the match stage is to find pairs of person-publication records that refer to the same person. If two people have the same or very similar names and also the same or very similar values on some combination of other attributes, i.e., collaborator, affiliation, location, content, then the two records probably refer to the same person, i.e., nascent person identifier. Bibliographic data typically include unique lists for publications but not for the people who wrote them. So, clustering person-publication records on a large scale requires an algorithm that can match records where names are the same or very similar where there is support from some combination of attributes with the same or very similar values and discriminate between records where names are the same or very similar but where there is *no* support from a combination of attributes. Disambiguation algorithms calculate similarity scores to classify records into clusters and, then, match records that exceed some threshold. Similarity scores for pairs of records are based on a combination of attributes with a weight assigned to each attribute, if an attribute is the same or very similar between records. Weights and similarity scores typically use a probability or odds-ratio metric, which is the default output from binary classifier type algorithms and relatively easy to interpret. A combination of attributes and weights is then input to the algorithm to compute the sum of the weights, i.e., the similarity score, for two person-publication records. Then, if the similarity score is above the threshold, then a match is declared.

The combination of attributes, weights, and thresholds, can then be fine-tuned to eliminate false-positives using a hand-curated training sample. The best fitting combination of attributes and weights is carried forward as an input for the algorithm to classify previously unseen records. Fine-tuning algorithms using hand-curated training samples have a number of known problems (Hastie et al., 2009). First, a linear weighted combination of attributes fails to capture interactions between their similarities. For example, if two person-publication records have the same or very similar name and the same or very similar affiliation, but the organization is large and has publications across many technical domains, then the relation between name and affiliation is probably conditional on content. Second, hand-curated training samples to find combinations and weights for attributes are typically a small and biased subset of the population (Li et al., 2014, p. 944). Finally there are risks of overfitting the predictive model to the training sample, which implies excess bias and variance when the predictive model is applied to new and previously unseen records (Hastie et al., 2009).

The machine learning literature (Li et al., 2014, pp. 944–945) helps to manage bias and variance while fitting predictive models to a training sample to find a combination of attributes, weights, and thresholds, that produces sensible results from the algorithm:

1) Train a probabilistic model: a) assumes only multidimensional order and thus captures non-linear and interaction effects among attributes, b) uses principles of probability theory to correct for transitivity violations among person-publication records, and c) classifies records into clusters using a likelihood-based framework.
2) Train model with large, diverse, and automatically generated training samples with high probability of matches and non-matches, which are drawn from across the entire population so that selection bias, training variance, and manual effort are reduced.
3) Use intentionally generic attribute categories so that the trained model can be applied to new and previously unseen data.

Viewed in this way the name disambiguation problem boils down to a classification exercise where the objective is to label pairs of person-publication records as *matches* or *non-matches*. A classification algorithm takes in a set of attributes associated with an object and, based on a set of previously “learned” representative examples, the algorithm uses these attributes to label the object with a class (Bishop, 2006). Here, the objects are pairs of person-publication records and the attributes are similarity scores obtained by comparing specific attribute fields contained in each record. For example, if two records contain the same or very similar names, then the similarity scores for these records is calculated from the similarity of other attribute fields in the record, e.g., collaborator, affiliation, location, content, and the class is either a match or a non-match. Nascent person identifiers can be constructed using a binary classifier with iterative clustering, i.e., filtering, for pairs of person-publication records to determine whether the two records refer to the same person, i.e., a match or a non-match.

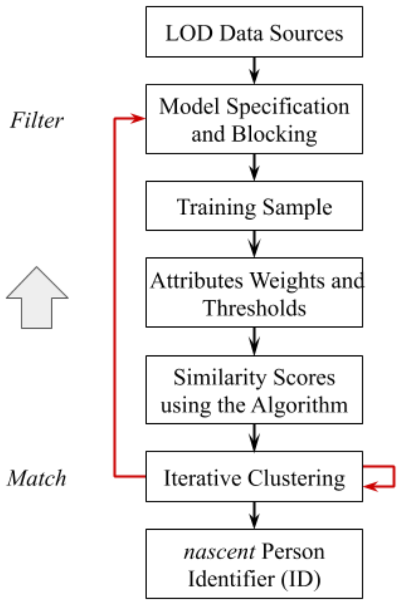

LOD Data Sources
Filter
Model Specification and Blocking
Training Sample
Attributes Weights and Thresholds
Similarity Scores using the Algorithm
Match
Iterative Clustering
nascent Person Identifier (ID)

Figure 12. The steps for a general name disambiguation procedure

The figure summarises the steps for a general name disambiguation procedure to identify nascent person identifiers from incomplete bibliographic data that includes several iterated steps. Based on Li and colleagues (2014, p. 945).

Figure 12 illustrates steps for a general name disambiguation procedure to identify nascent person identifiers from incomplete bibliographic data that includes several iterated steps. The first step is to specify an initial model for person attributes to represent person-publication pairs, i.e., select attributes to compare and define how to compare records using the algorithm. Once we have settled on an initial model, we draw one or more training samples of person-publication records from the population where the class labels are “known” from manual annotation of the records. We use training samples to learn the probability of pairs of records being declared a match using similarity scores from the initial model. We then use the model to compute similarity scores for another training sample and again evaluate whether the scores provide enough evidence to declare a probable match on nascent person identifiers for a pair of records. Finally, any conflicts are resolved in the match and non-match classifications by manual annotation and a corrected cluster of publications referring to one or more nascent person identifiers can be carried forward into a ‘golden’ dataset. This approach can be used to fine-tune the combination of attributes, weights, and thresholds in the initial model, which is based on previous research and theory. Later, we progress to a less naive model, which is better adapted to the specific empirical context for the data at hand. However, the application of this approach to every pair of records in the population has a prohibitive computational cost.

Blocking records is a typical heuristic used to reduce computational effort for binary classification problems. Specifically, we organize records into clusters, e.g., SSM on family name and given name, where nascent person identifier matches are most likely to be revealed by the algorithm. Thus, we only have to compute similarity scores for pairs of records within a block, as opposed to all records, which represents a major saving. We iteratively develop working clusters of person-publication records, which we accumulate in entity cards, using similarity scores based on the algorithm and blocking rules. An entity card contains one or more person-publication records assigned to a nascent person identifier. After several rounds of agglomerative clustering, a nascent person identifier may be assigned several entity cards. Thus, the person-publication record, entity card, and nascent person identifier are related in a sequence of many to one relations that in conjunction with the general procedure, illustrated in Figure 12, allows us to address the name disambiguation problem. Repeated rounds of agglomerative clustering continue until there are no more records in the block that could be matched to existing clusters. Li and colleagues (2014) describe this as a semi-supervised classification approach. Figure 12 provides an overview of our version of the procedure including, filtering records using model specification and blocking, drawing training samples, choosing attribute weights and thresholds, calculating similarity scores using the algorithm, finding record matches using iterative clustering. Our contribution is to recast the name disambiguation approach in an RDBMS harmonization framework that preserves more of the existing details from the LOD sources, i.e., the “curse of dimensionality” (Hastie et al., 2009, p. 22), compared to the problems of machine learning approaches with high-dimensional nonlinear functions.

Filter

The principal objective of the *filter* stage is to reduce the number of false positives, i.e., lumping, from the match stage. The filter stage finds and rejects false positive matches for person-publication records that do not refer to the same person. If two people have the same or very similar names but a different, albeit similar, values because we change the combination of attributes in the model, i.e., collaborator, affiliation, location, content, or the blocking rule to cluster records, then the two records probably do not refer to the same person after all. Thus we use alternative attribute combinations and blocking rules, to check the robustness of matches from previous rounds of agglomerative clustering.

In principle, we would like to classify every pair of person-publication records from the combined LOD sources as a match or non-match. However, exhaustive pairwise comparison requires quadratic runtime and for 44.8 million person-publication records in the combined bibliographic LOD sources would require 2 quadrillion comparisons, i.e., two followed by 15 zeros, making the full problem computationally infeasible. As suggested above, blocking is an approach to making pairwise comparisons computationally feasible:

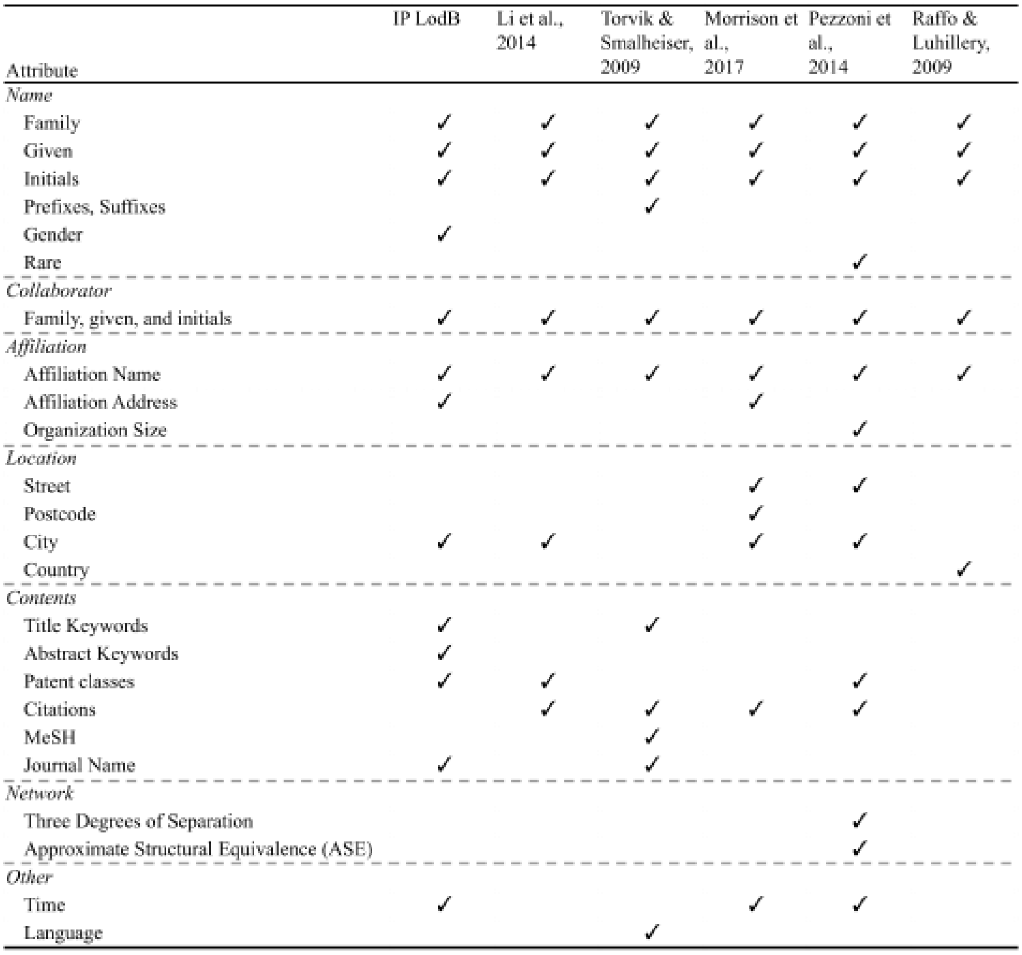
1) Block the records by clustering into smaller groups, e.g., same family name and given name.
2) Restrict pairwise comparisons between records within those blocks.

The blocking heuristic is applied by clustering person-publication records into groups that are likely to contain true matches and possibly only true matches. The person-publication records input to the algorithm have been augmented and validated from several secondary sources, as shown in person tables in Figures 6 and 9, to ameliorate missing data in the LOD source records. Choosing a combination of attributes for a blocking rule is a difficult tradeoff. On one hand, if the blocks are too big, then they contain too many records to reduce quadratic runtime. On the other hand, if the blocks are too small, then they contain too few records to detect false negatives, i.e., splitting errors, present in the records from both LOD sources. The tradeoff between runtime and accuracy can be managed by iteratively refining the blocking rules.

Iterative blocking rules help to manage the computational load by first inputting records to the algorithm using relatively fine grained rules to reduce the surface area of the problem and, only later, inputting records using relatively coarse grained rules. Early iterations use relatively fine grained blocking rules, i.e., complete text strings for family and given names, e.g., Duck, Donald. Once we use the algorithm to compute working clusters of person-publication records, i.e., entity cards, based on the blocking rule and manually validating the matched records, we can reduce the effective size of the database by collapsing person-publication records together into entity cards and nascent person identifiers. Subsequently, coarser grained blocking rules become more feasible, i.e., incomplete text strings for family names and given names represented by initials, e.g., Duck, D.F. Iterative blocking rules allow the algorithm to scale and evaluate a larger selection of potential matches than any single blocking rule would allow (Li et al., 2014). Thus, moving from restrictive blocking rules to permissive can reveal person-publication record matches that might otherwise be missed, as illustrated in the person table in Figure 9.

The purpose of iterative blocking is to reduce the number of false negatives left behind after a single iteration, e.g., only SSMs on complete family and given names, which fail to capture initials, common name variants, e.g., William vs Bill, minor misspelling, and inversions. As the number of records in a block increases, the probability of finding a match reduces. Taken together, data preparation, iterative blocking and algorithm produces a list of hand-crafted rules that *inter alia* preserves the fidelity of the input data used for prediction and to reduces the number of ‘surprising’ matches that machine learning approaches can produce. Nevertheless, despite achieving a compromise between scale and performance, our general disambiguation procedure remains an imperfect heuristic. Despite our best efforts, false positive matches between person-publication records can still occur. Furthermore, using agglomerative clustering through entity cards, lumping errors can cascade from one round to the next. A fixed threshold tends to further compound cascading false positives but on balance, the risk of lumping errors from a fixed threshold is outweighed by decreased false negatives, i.e., splitting errors, and scalability gains from an iterative approach (Li et al., 2014, p. 946).

Table 3. Attribute comparison for name disambiguation across studies



Attribute	IP LodB	Li et al., 2014	Torvik & Smalheiser, 2009	Morrison et al., 2017	Pezzoni et al., 2014	Raffo & Luhillery, 2009
Name						
Family	✓	✓	✓	✓	✓	✓
Given	✓	✓	✓	✓	✓	✓
Initials	✓	✓	✓	✓	✓	✓
Prefixes, Suffixes			✓			
Gender	✓					
Rare					✓	
Collaborator						
Family, given, and initials	✓	✓	✓	✓	✓	✓
Affiliation						
Affiliation Name	✓	✓	✓	✓	✓	✓
Affiliation Address	✓			✓		
Organization Size					✓	
Location						
Street				✓	✓	
Postcode				✓		
City	✓	✓		✓	✓	
Country						✓
Contents						
Title Keywords	✓		✓			
Abstract Keywords	✓					
Patent classes	✓	✓			✓	
Citations		✓	✓	✓	✓	
MeSH			✓			
Journal Name	✓		✓			
Network						
Three Degrees of Separation					✓	
Approximate Structural Equivalence (ASE)					✓	
Other						
Time	✓			✓	✓	
Language			✓			

The table summarises attributes used in name disambiguation processes to identify nascent person identifiers from incomplete bibliographic data. As before, attributes are organised into names and four other generic categories, i.e., collaboration, affiliation, location, and content. However, here we also include some additional attributes used in other studies, some of which we may include in future studies, e.g., citations. 'Patent classes' include the CPC taxonomy. MeSH = medical subject headings. Based on Morrison and colleagues (2017, p. 3).

Table 3 lists name and other attributes used in our study and compares them with those used in other similar name disambiguating studies using bibliographic data. We organized the Table according to our clean and parse procedures for name text strings and our four generic categories for other attributes, i.e., collaboration, affiliation, location, and content. Table 3 shows two things: 1) that attributes used in previous studies are distributed across our generic categories in a reasonably balanced way and 2) our selection of other attributes compare favorably in scope to those used in previous studies.

Method

Overview

The main sources of data for our work are EP LOD and SN LOD databases. While they contain a lot of information we are using to compile attribute data, there are some additional data available in other datasets, which essentially enhance information about considered entities. We thus also use other databases, such as the PATSTAT database and the Global Research Identifier Database (GRID). We also use various collections of simple data, such as country abbreviation list, geonames list, and list of gender assignments to given names per country. We use these to extract partial data from the fields containing several partial information. For example, from affiliation addresses, we extract cities and countries, from the given name, we extract gender etc. Extraction of additional information is done also in a more sophisticated way, such as using an adapted RAKE algorithm for keywords. We are using publication abstracts and titles to extract relevant keywords and assign them to the entities in consideration.

Data retrieval sources[17]

Data for several attributes is extracted from **Patstat** - the core data product by the EPO which contains data on patents. We use only Patstat (2019). Interconnection of databases is straightforward, since EP LOD and PATSTAT contain links between corresponding patent applications. We have thus imported several variations of the inventors names (Person name (PERSON_NAME), Harmonized Applicant Name from OECD disambiguation (HAN_NAME), person name which have been harmonised according to the Univ. Leuven harmonization procedure (PSN_NAME), and person name in original language (PERSON_NAME_ORIG_LG). We also added their addresses (Person address (PERSON_ADDRESS), their NUTS code (NUTS) and country (person_ctry_code) and additional information on their field of work, by bringing in also any potential additional information on CPC classification symbols (CPC_CLASS_SYMBOL) - and added them to the information available directly from the EPO LOD. Lastly, since the bulk download files from EPO LOD do not include abstracts (for our keyword extraction), we have also added the patent abstracts from the field abstract of application (APPLN_ABSTRACT).

The **Global Research Identifier Database (GRID)**[18] uses uniform resource identifiers (URIs), hence it is another linked open database. GRID includes information on almost 100,000 organizations, out of which about 30% are companies, 20% are higher education institutions (HEI), with about 10% nonprofit and 10% hospitals. The database includes several variables, such as address, type, the URl of the organization etc. (GRID, 2020). The use of GRID is accommodated by the fact that SN LOD contains affiliation links with the records on

[17] Note that the list does not include the main two data sources (i.e. our main axis), but we describe them in various other places (such as in Deliverable 2.1 and on iplod.io webpage under Aligning-->Data Sources).
[18] Openly available at: https://www.grid.ac/

organizations in GRID. We have connected 34,676,400 records, that include information on organizations, with the organizations' corresponding GRID record and provide working links.

GeoNames[19] is a freely available geographical database covering all countries and containing over eleven million place names that are available for download. They provide URI for their data, but also have a data dump available, from where we used the information available from cities5000.zip file. Currently, the data is used for disambiguation purposes and we do not provide a working link between person addresses (not even for registered users) and the Geonames data, due to potential General Data Protection Regulation (GDPR) restrictions.

ISO-3166-Countries-with-Regional-Codes[20] is a country abbreviation list, with the data dump available in the .csv format on the GitHub repository.

The **World Gender-Name Dictionary (WGND)**[21] compiles the information from 13 different sources (from either national public institutions and previous gender studies and including some limited manual check) on gender attribution for first names. Combined this dataset covers over 173 different countries and includes 6.2 million names for 182 different countries disambiguating the names for PCT inventors (Martinez et al, 2016). We apply the gender in connection to country (using their table wgnd_langctry, available in .exe format, which provides a more limited data), due to specific names that could indicate different genders in different countries, e.g., Andrea in Italy and Spain. Due to our more restrictive approach (we use this as a part of the blocking), we have attributed gender to approximately 55% of SN person-records and a little over 60% to EPO person-records.

The **Crossref**[22] contains over 120 million records and is one of the major sources of scholarly data for publishers, authors, librarians, funders, and researchers. The metadata set consists of 13 content types, including not only traditional types, such as journals and conference papers, but also data sets, reports, preprints, peer reviews, and grants. The metadata is available through a number of APIs, including REST API and OAI-PMH. The metadata is not limited to basic publication metadata. It can also include e.g. abstracts and links to full text, funding and license information, citation links, and the information about corrections, updates, retractions (Hendricks et al, 2020).

Data Pre-processing

1. Family and given names were split; where letters "c/o" appeared, the applicant was also extracted.
2. Gender was extracted from given names using World Gender Name Dictionary.

[19] Openly available at: https://www.geonames.org/ with data dump available at: http://download.geonames.org/export/dump/

[20] Downloadable at: https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes/blob/master/all/all.csv

[21] Freely available in Database repository at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YPRQH8

[22] API information available at: https://www.crossref.org/education/retrieve-metadata/rest-api/

3. Countries and cities were extracted from institution data wherever possible using GeoNames database
4. Keywords were extracted from patent and publication titles and abstracts with the RAKE algorithm

Our algorithm for extracting **countries and cities** from institution from SN LOD, where those appeared is described below:

1. The institution string was matched with lists of countries from the GeoNames database. If a country name appeared, it was chosen as the correct country for the institution. Countries were compared in the following order: full country names longer than 5 letters, full country names between 3 and 4 letters, country abbreviations, country code.
2. For a chosen country the list of all cities of population over 5000 was collected from the GeoNames database. The institution string was matched with city names in the following order: city names over length of 5, alternative city spellings over length of 5, city names under length of 5, alternative city spellings under length of 5.

Example 1
Affiliation: Departments of Biochemistry, Surgery, and Statistics, R.N.T. Medical College, Udaipur, India
Extracted Country Code: IN
Extracted City: Udaipur

Example 2
Affiliation: Institut für Allgemeine Physik, Technische Universität Wien, Wiedner Hauptstr. 8-10, Wien, A-1040, Austria
Extracted Country Code: AT
Extracted City: Vienna

Analysis

Our method for disambiguating people builds on most of the successful approaches described above. In short, the method is executed by the following algorithm:

1) Determine initial block.
2) Create an entity card for every block entry.
3) **Repeat.**
   a) Compute similarity for each pair of entries.
   b) Join the pair with the highest similarity in a new entry and remove the initial two entries.
4) **Until** no pair is joined.

Below we explain each of the steps in more detail.

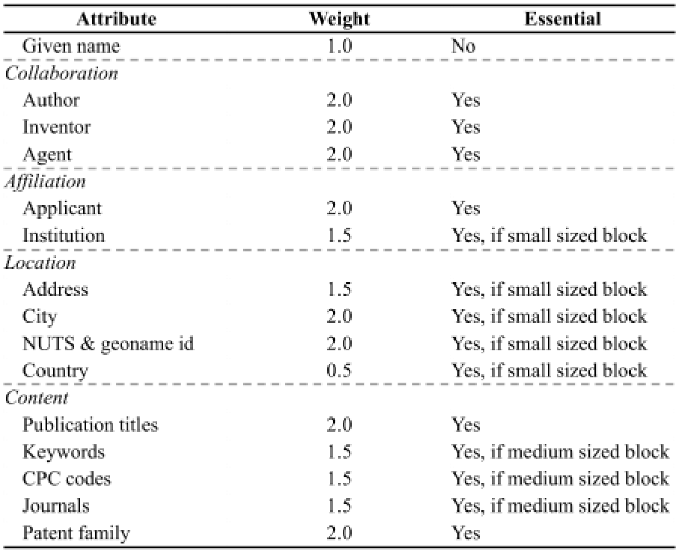
Determine initial block. Our blocking is based on family names. In particular, we collect in one block all persons having the same cleaned text string for family name. Note that blocking in our method can be extended in an iterative process. Namely, after blocks of persons with the same family name are disambiguated, and so the number of persons is essentially reduced, a blocking can be, e.g., performed on the affiliating institutions for catching persons with changed names or spelling errors.

Create an entity card. Entity cards are collections of attribute values obtained for an entry. The majority of data is usually extracted from person-publication tables; however, the extraction is not limited to that. For example, SN LOD provides a dataset of persons' data retrieved from various sources, containing, e.g., information about affiliations and ORCID ids. Apart from the basic attributes: family name, given name (including middle names), gender, and activity period, we collect a number of other attributes, which we divide into four categories: collaboration, affiliations, location, i.e. either affiliation address or inventor address, and content, i.e. technical domain. In the collaboration category, we collect lists of authors and inventors, i.e., co-inventors of patent documents, co-authors of scientific publications. In the affiliation category, we include lists of applicants and institutions, i.e., applicants and agents on patent applications. In the location category, we collect addresses (address of the affiliating institution or any other address assigned to the person within available data). From addresses, we extract cities and countries to infer possible matches on a larger scale. In some cases, NUTS and Geoname identifiers are available, helping us to match two attributes in an easier way. In the content category, we collect attributes related to work areas of persons. We collect titles and abstracts of their publications, keywords (which are extracted from the former), journals in which they publish, patent families of the patent applications, and fields of research (e.g., CPC codes assigned to patent applications).

Compute similarity. For a pair of entries, we need to determine if they represent the same person and we do that by comparing their entity cards. We take a very conservative approach in the sense that we match two persons only if we are very confident that they represent the same entity. The similarity score S for a pair of entries is determined in three steps. First, we check if assigned genders of both entries are different, in which case we return $S=0$. Next, we compute the similarities between given names of both entries. Since our method is iterative, every entry can have a list of multiple given names (e.g., {Jon; J.; Jonathan}) and we compute similarities for all pairs (a, b), where a is a given name of the first person and b is a given name of the second person. If the similarity of **all** pairs is higher than a given threshold, we proceed to the third step; otherwise, we return $S=0$.

Next, we proceed with the actual computation of the similarity score. In a similar manner as for the given name, we compute similarities for all attributes that two entries have in common. This is *mainly* done by comparing single values in attribute data, e.g. coauthor names, and searching for matching data. If there is an exact match within attribute data for both entries, we set the attribute similarity to 1. If the match is not complete, but still can be detected, e.g., address matches in the street, we assign a value according to the degree of similarity.

Table 4. A table of attributes used by our method, their weights, and essentiality.



Attribute	Weight	Essential
Given name	1.0	No
Collaboration		
Author	2.0	Yes
Inventor	2.0	Yes
Agent	2.0	Yes
Affiliation		
Applicant	2.0	Yes
Institution	1.5	Yes, if small sized block
Location		
Address	1.5	Yes, if small sized block
City	2.0	Yes, if small sized block
NUTS & geoname id	2.0	Yes, if small sized block
Country	0.5	Yes, if small sized block
Content		
Publication titles	2.0	Yes
Keywords	1.5	Yes, if medium sized block
CPC codes	1.5	Yes, if medium sized block
Journals	1.5	Yes, if medium sized block
Patent family	2.0	Yes

Table 4 specifies weights used for each attribute in the algorithm and some essential conditions where the similarity score is bypassed in favor of a discrete match or non-match decision. Clearly, not all attributes have the same importance for the matching process, and therefore, we assign importance weights to each of them. Also, if a matching attribute strongly indicates that two entries are the same person, e.g., having the same coauthor, we denote the attribute as *essential*, and we use it to override the value of the similarity score as described in the next paragraph. For some attributes, we decide if they are essential regarding the size of a current block (small sized is a block with at most 200 entries and medium sized is a block with at most 1000 entries). For example, it is more probable that in the same city, there are two persons with the same family and given name if there are a lot of persons with such a name. In the current version, the importance weights are assigned based on experience gained by manual disambiguation. In the future, we plan to assign their values based on machine learning algorithm's evaluation.

When all attribute similarities for an attribute category are computed, we determine an attribute with the highest similarity S_{category} and use it as a representative of the category when computing the similarity score S. We also use the attribute's weight as the weight W_{category} for the category. S is then computed as

$$S = \frac{(S_{\text{give}} \times W_{give}) + (S_{coll} \times W_{coll}) + (S_{affi} \times W_{affi}) + (S_{loca} \times W_{loca}) + (S_{cont} \times W_{cont})}{W_{give} + W_{coll} + W_{affi} + W_{loca} + W_{cont}}$$

where S = the similarity score for a pair of entity cards. The categories are abbreviated as follows: give = given name, coll = collaborator, affi = affiliation, loca = location, and cont = content. If S is higher than a given threshold, we join two entries in one.[23]. If not, we check if

[23] Currently, in version 110, based on our analysis, the threshold is set to 0.6 and the affiliation is merged with the address category, to avoid false positives. Thus the algorithm is: *S = (Sgive * Wgive + Scoll * Wcoll + Sloca * Wloca + Scont * Wcont)/(Wgiv + Wcoll + Wloca + Wcont)*

there was some high similarity within any essential attribute and, if affirmative, we set *S=1*, to force a match.

Results

To validate our results we have built a golden dataset (i.e., a training set), which is connected to our version 110 (also the version available online at search query at iplod.io). This is a hand curated set, generated by members of the IPLOD group themselves. The records were validated by using internal data, as well as cross-checked using external data sources. This golden dataset contains a record for each pair of initial entities that are referring to the same person. All together we checked 2,895 distinct original entities (i.e. persons before disambiguation), which together give 1,111,771 pairs of original entities that have been manually confirmed to represent a match. We now turn to some examples within this golden dataset.

Auditing Initial Results and Manual Annotation

The following are three examples that show how our name disambiguation procedure works using data from EPO and SN LOD sources, again after our best efforts to augment and validate primary data from secondary sources. The three examples show the source text strings and how we use those data to make inferences about whether person-publication records refer to the same or different people.

Example 1. Match based on Name and Affiliation

Family name: Vučković **Given name:** Nancy **Gender:** Female **Activity period:** from 2002 to 2002 **Database:** SN LOD	**Family name:** Vučković **Given name:** Nancy **Gender:** Female **Activity period:** from 2013 to 2013 **Database:** EP LOD
Collaboration **Author:** Green, Carla A.	Collaboration **Inventors:** Crawford, Richard P.; Han-Steen-Izora, Muki; Morris, Margaret E.
Affiliation **Institution Name:** Kaiser Permanente Center for Health Research, Oregon, Portland; Intel, United States **Grid:** 414876.8; 419318.6	Affiliation **Applicant Name:** Intel Corporation **Agent Name:** Jennings, Vincent Louis **Grid:** 419318.6
Location **Institution address:** Kaiser Permanente Center for Health Research, Oregon, Portland; Intel, United States **Cities:** Tucson; Oakland, Portland, Santa Clara **Country:** United States **Geoname:** 5318313, 5378538, 5746545, 5393015	Location **Inventor Address:** 20th Ave, Portland, Oregon 97202 **City:** Portland **Country:** United States of America **Geoname:** 4650946
Content **Publications:** Adapting to psychiatric disability and needs for home- and community-based care **Journal:** Mental Health Services Research	Content **Publications:** Devices, systems, and methods for enriching communications

Similarity Score

$$S = \frac{(1 \times 1) + (0 \times 1) + (1 \times 1.5) + (1 \times 1) + (0 \times 1)}{1 + 1 + 1 + 1 + 1} = \frac{3.5}{5} = 0.7$$

Example 1 shows a perfect match (similarity 1) in the given name and in affiliation. Matching affiliation has weight 1.5, all the other weights in this case are 1. Note that although there is a match also between cities (Portland), we only choose one attribute (the one with the highest value) as a representative within the location category. If we used a 0.6 threshold for matching pairs, this pair would be matched by the computed similarity score. Alternatively, since the affiliation attribute is essential, we reset S to 1 and consequently match the two entries.

Example 2. Non-Match based on Name and Location

Family name: Vučković **Given name:** I. **Gender:** Unknown **Activity period:** from 1993 to 1993 **Database:** SN LOD	**Family name:** Vučković **Given name:** Ivan **Gender:** Male **Activity period:** from 2017 to 2017 **Database:** SN LOD
Collaboration **Authors:** Hrzenjak, M.; Ilić, Z.; Jurin, M.; Kis, S.; Stipančić, I.; Žrković, N.	Collaboration **Authors:** Balint, M.; Kisfali, P.; Kučinić, M.; Mikes, T.; Pauls, S. U.; Szalontai, B.; Szivak, I.; Vadkerti, E.
Affiliation **Institution Name:** Military hospital, Zagreb **Grid:** -	Affiliation **Institution Name:** Civil and architectural dept., Zagreb **Grid:** -
Location **Institution Address:** Military hospital, Zagreb **City:** Zagreb **Country:** Croatia **Geoname:** 3186886	Location **Institution Address:** Civil and architectural dept., Zagreb **City:** Zagreb **Country:** Croatia **Geoname:** 3186886
Content **Publications:** Host defense dysfunction in trauma, shock and sepsis; The influence of liver regeneration on skin wound healing and lymphocyte growth attributes	Content **Publications:** Ecological divergence of *chaetopteryx rugulosa* species complex linked to climatic niche diversification **Journal:** Hydrobiologia

Similarity Score

$$S = \frac{(0.8 \times 1) + (0 \times 1) + (0 \times 1) + (1 \times 2) + (0 \times 1)}{1 + 1 + 1 + 2 + 1} = \frac{2.8}{6} = 0.47$$

Example 2 shows a relative match (similarity 0.8) in the given name and in a perfect match in the city (similarity 1.0). Matching city has weight 2.0, all the other weights in this case are 1. If we use a 0.6 threshold for matching pairs, this pair is not matched by the computed similarity score. However, since the city attribute is essential whenever we have a relatively small block of persons (at most 200), we reset S to 1 and consequently match the two entries. Unfortunately, in this case, the algorithm's decision is incorrect and we obtain a false positive.

Example 3. Match based on Name and Location

Family name: Vučković **Given name:** Slavica **Gender:** Female **Activity period:** from 1997 to 2016 **Database:** SN LOD	**Family name:** Vučković **Given name:** Slavica **Gender:** Female **Activity period:** from 2000 to 2000 **Database:** EP LOD
Collaboration **Authors:** Auditore-Hargreaves, K.; Avery, J.; Butler, J.; Catley, l.; Clark, G. J.; Connor, T.; Curley, C.; …; Hart, D. N. J.; …	Collaboration **Inventor:** Hart, D. N. J.
Affiliation **Affiliation:** School of medicine, University of Queensland, 4072, Herston, QLD; Mater medical research institute, Raymond Terrace, 4101 South Brisbane, Queensland; Royal prince Alfred hospital; Christchurch hospital, Christchurch;	Affiliation **Affiliation:** The corporation of the trustees of the order of the sisters of mercy in Queensland **Agent:** Jones, E. L. **Applicant:** The corporation of the trustees of the order of the sisters of mercy in Queensland
Location **Affiliation:** School of medicine, University of Queensland, 4072, Herston, QLD; Mater medical research institute, Raymond Terrace, 4101 South Brisbane, Queensland; Royal prince Alfred hospital; Christchurch hospital, Christchurch; … **City:** Swan View; South Brisbane; Christchurch; Brisbane **Geoname ID:** 2060771 **Countries:** Australia; New Zealand	Location **Affiliation:** The corporation of the trustees of the order of the sisters of mercy in Queensland **City:** Swan View **Geoname ID:** 2060771 **Country:** Australia
Content **Publications:** Dendritic cell surface molecules; Dendritic cells in fundamental and clinical immunology; …; Immune responses in multiple myeloma: role of the natural immune surveillance and potential of immunotherapies; … **Keywords:** dendritic cells; inflammatory bowel disease; macrophages; multiple myeloma; myeloma cells	Content **Publications:** Dendritic cell-specific antibodies **CPC codes:** A61; C07

Similarity Score

$$S = \frac{(1 \times 1) + (1 \times 2) + (0 \times 1) + (1 \times 2) + (0 \times 1)}{1 + 2 + 1 + 2 + 1} = \frac{5}{7} = 0.71$$

Example 3 shows a perfect match (similarity 1.0) in the given name, collaborator name, and in the geoname id (also in the city, but geoname id is narrower). Matching collaborator and geoname id have weight 2.0, the other two weights in this case are 1. Since we are using threshold 0.6 for matching pairs, this pair is matched by the computed similarity score.

One of the persons inside our golden dataset is Ferrarese Carlo, an IPLOD unique person (IP Lodb ID: ac987ba2-ae03-4212-912a-b7179f9392b7) containing both SN LOD records and an EP LOD record.[24] Hand checking showed that we have indeed disambiguated correctly (hence we detected no false positive nor any false negatives). This is somewhat an extreme case in terms of the volume (but not complexity) where we merged hundreds of SN LOD persons inside this IPLOD unique person, and connected them to the corresponding EPO LOD person. There were 3 types of records. First are individual records for numerous encyclopedia entries (within the Encyclopedia of psychopharmacology), the second type is one person but with many publications attributed to them (previously disambiguated by SN LOD). The third is a patent in the field this person is active in. There is overlap in the field, as well as common co-authors / co-inventors, whereas the affiliation, albeit same, is not displayed directly in SN LOD, but we needed to check the contributors list via links provided by SN LOD – showing that the promise of linked data leading to more discovered data can indeed be fulfilled.

In Table 5, we consider Lamprecht, Björn.[25] This is a connected person that includes three person-records: Lamprecht B., an SN person, and two EPO persons, both named Lamprecht, Björn. When we observe their respective entity cards we can see that our algorithm took advantage of various attributes that cross these person-records. We can observe their collaboration (collaborators and organizations they collaborate with (i.e. affiliations and applicants, respectively). First we observe collaborators (social network) with several matches, including exact matches on coauthors last and first names.

Table 5. Collaborators comparison (Lamprecht, Björn)

Collaborators		Affiliation
SN LOD	EPO LOD	SN LOD
	Coinventors: bonifer, constanze dorken, bernd lamprecht, bjorn mathas, stephan	Affiliations: ● German Cancer Research Center, Heidelberg, Germany ● Hematology, Oncology and Tumorimmunology, Max Delbrück Center for Molecular Medicine, Berlin, Germany Hematology, Oncology and Tumorimmunology, Charité, University Hospital Berlin, Campus Virchow Klinikum, Berlin, Germany ● Max-Delbrück-Center for Molecular Medicine, Berlin, Germany. Hematology, Oncology and Tumor Immunology, Charité–Universitätsmedizin Berlin, CVK, Berlin, Germany. ● Max Delbrück Center for Molecular Medicine, Berlin, Germany
		EPO LOD

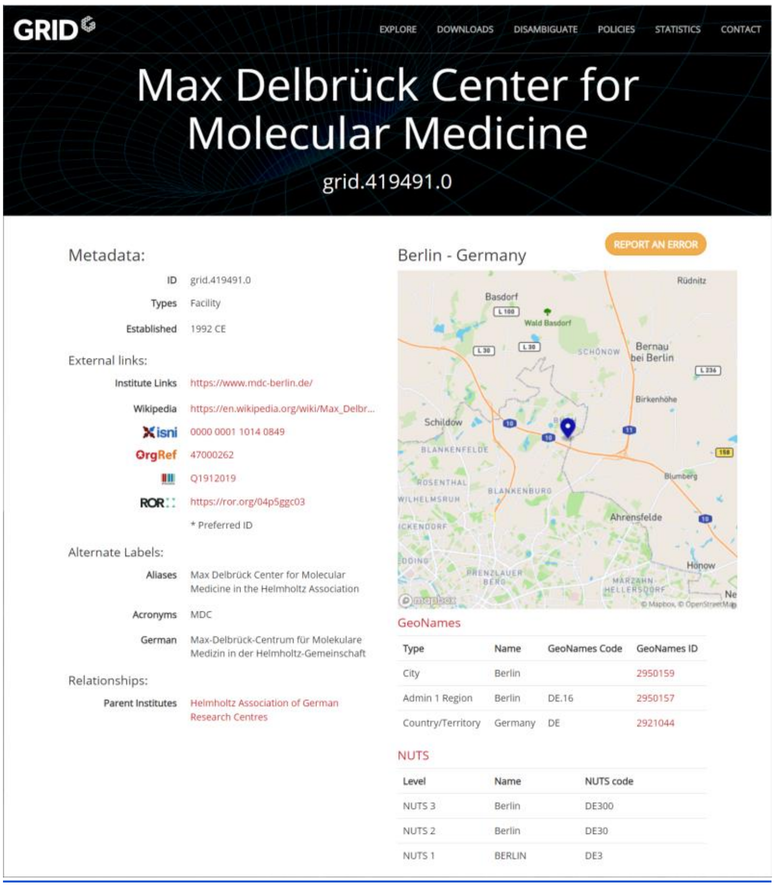
[24] The numbers apply to the version 110, which is available currently for non-registered users, but with all versions being available at any point to registered users also through the iplod.io search query.

[25] IPLodB ID: 238e03d6-3ef1-4d05-9311-ebf439123ec0

Coauthors:	Collaborates with Applicants:
anagnostopoulos, i barlow, rachael begay, valerie bonifer, c bouhlel, mohamed amine callen, d. f. cockerill, p n dorken, bernd gatjen, marcel gerlach, kerstin giefing, m gloger, marleen heinig, kristina hopken, u e hummel, m janz, m johrens, k jundt, f kitagawa, m kochert, k kreher, s kumar, raman lenze, d leutz, a lipp, martin mathas, s mensen, angela rehm, armin richter, j siebert, r soler, eric stadhouders, ralph stein, h ullrich, k walter, korden wurster, k d zimber-strobl, ursula	● Max-Delbrück-Centrum für Molekulare Medizin in der Helmholtz-Gemeinschaft ID: dc25efe0-1fe4-c68f-29e2-420207fe748e ● Max-Delbrück-Centrum für Molekulare Medizin Berlin-Buch ID: 03c976e5-e41f-f769-aaf9-a862e66909bc

On the left hand side, the table displays the corresponding SN LOD and EPO LOD person’s collaborators, co-authors for SN and co-inventors for EP LOD, showing that almost all coinventors also appear as coauthors inside publications produced by our IPLOD person (in green text). On the right hand side the corresponding affiliations are also highlighted.

We turn then to affiliation, and we can also observe that for example Max Delbrück Centrum, there is a hyperlink established leading to the corresponding GRID record.



Max Delbrück Center for Molecular Medicine

grid.419491.0

Metadata:

ID	grid.419491.0
Types	Facility
Established	1992 CE

External links:

Institute Links	https://www.mdc-berlin.de/
Wikipedia	https://en.wikipedia.org/wiki/Max_Delbr...
isni	0000 0001 1014 0849
OrgRef	47000262
	Q1912019
ROR	https://ror.org/04p5ggc03
	* Preferred ID

Alternate Labels:

Aliases	Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Acronyms	MDC
German	Max-Delbrück-Centrum für Molekulare Medizin in der Helmholtz-Gemeinschaft

Relationships:

Parent Institutes	Helmholtz Association of German Research Centres

Berlin - Germany

GeoNames

Type	Name	GeoNames Code	GeoNames ID
City	Berlin		2950159
Admin 1 Region	Berlin	DE.16	2950157
Country/Territory	Germany	DE	2921044

NUTS

Level	Name	NUTS code
NUTS 3	Berlin	DE300
NUTS 2	Berlin	DE30
NUTS 1	BERLIN	DE3

Figure 13. GRID link connected to IPLOD organization (Lamprecht, Björn)

The URI (https://www.grid.ac/institutes/grid.419491.0) of the affiliation (Max Delbrück Centrum) is reachable through an active link on the iplod.io webpage.

Figure 13 shows how to take advantage of institutions in SN LOD and applicants in EPO LOD information. However, the affiliation names here have several name variations, from "Max-Delbrück-Centrum" to "Max Delbrück Centrum". Furthermore, language variation can occur in organization names, e.g. the IPLOD person Bertagnoli, Stéphane,[26] which contains both EPO and SN persons, and applicant "Ecole Nationale Veterinaire Toulouse", but has under affiliations also the English correspondence of the original name, the "National Veterinary School of Toulouse, University of Toulouse". Besides having aggregated several versions of affiliations' language and different spelling variations for our IPLOD persons, the GRID connected records for the Max Delbrück Centrum and the link is available via iplod.io link.

[26] IPLodB ID: 5f8b4e60-2aeb-4148-9635-6db19ac3a377

Table 6. Location and content categories.

Countries: CROATIA	Countries: REPUBLIC OF CROATIA
Cities: Zagreb	Cities: zagreb
Geocodes: 3186886	Geocodes: 3186886
Publications: active deformation of the northern adriatic region:results from the crodyn geodynamical experiment the adria microplate: gps geodesy, tectonics and hazards	Publications: determination of horizontal and vertical movements of the adriatic microplate on the basis of gps measurements geodesy for planet earth

The boxes on the left and right display the corresponding city as well as geocodes. Underlined are the correspondences between the publication titles and the keywords therein.

Table 6 shows how IPLOD person Marjanovic, Marijan is disambiguated using location and content attributes.[27] The above person combines three original persons with different first names: “Marian”, “Marijan”, “M.”, respectively, and there are no common collaborators. These original persons are connected however *inter alia* through city (Zagreb) and geocodes (3186886), but not through the direct address data. We can also see that we catch the different spelling of countries (Croatia vs. Republic of Croatia), since we are using different geolocational solutions (form the country abbreviation list to others) Furthermore, the list of publications also shows that their technical domain is consistent.

Finally, Table 7 shows name disambiguation for Friedberg, Thomas Herbert,[28] which merges two SN LOD persons (both designated as Friedberg T.) and two EP LOD persons (both designated as Friedberg, Thomas Herbert). For content, beside comparing on CPC for EPO records (here we would find a match between the EP LOD persons) and on journal names for SN persons, we adapted the RAKE algorithm for inter-database matching (EPO LOD- SN LOD). (there would be no match in journal because the first SN person has listed as journals “Breast cancer research” and the other one “Archives of gynecology and obstetrics”, the “Journal of molecular medicine” and “Pharmaceutical research”. The Table shows an extract for the matching on one ‘essential’ keyword, i.e., cytochrome p450. In this example the SN LOD person Friedberg, T. (sg:person.0726356556.27) has ample keywords and the EPO person Friedberg, Thomas Herbert,[29] has limited keywords.

[27] IPLodB ID: ab26553c-e363-424d-961c-f2d2c0e7cce9

[28] IPLodB ID: 6d1c598f-cc42-4b9b-ab68-2472b7f78e14

[29] IPLodB ID: 69bd73cb-f031-6cc5-e10a-950c0910f0de

Table 7. Matching keywords in publications and patents

Keywords:	Keywords:
archives; bhk21 cells; biological monitoring; biological reactive intermediates; carcinogenic metabolites; cell lung cancer tissue prevalence; clinicopathological significance; cytochrome p450 protocols; die bedeutung der; doxepin isomers; drug metabolism pathways; drug metabolizing enzymes; environmental hygiene iii; expression fur die prognose beim ovarialkarzinom; functional human cytochrome p450 monooxygenase systems; generate new mutagenicity test systems; gynakologie und geburtshilfe 1992; heterologous microsomal epoxide hydrolase; mainly catalyzed; microsomal epoxide hydrolase membrane topology; molecular medicine; oxidative drug metabolizing enzymes; pharmaceutical research; several hepatic proteins related; stable expression; subcellular level; v79 chinese hamster cells;	cytochrome p450 expression;

Underlined are the matching keywords we generate with our adapted RAKE algorithm.

Reduction in unique persons pre- and post-disambiguation

We now turn from individual cases to shedding light on the disambiguation success rates. We have selected 100 randomized surnames and checked our disambiguation success rate. Within those were 37,832 original persons (i.e. persons before the disambiguation) from either EPO or SN LOD, which we then disambiguate, and end up with 13,530 unique persons. These persons are either connected within their original database (e.g., EPO person to EPO person) or between databases (EPO person to SN LOD person).

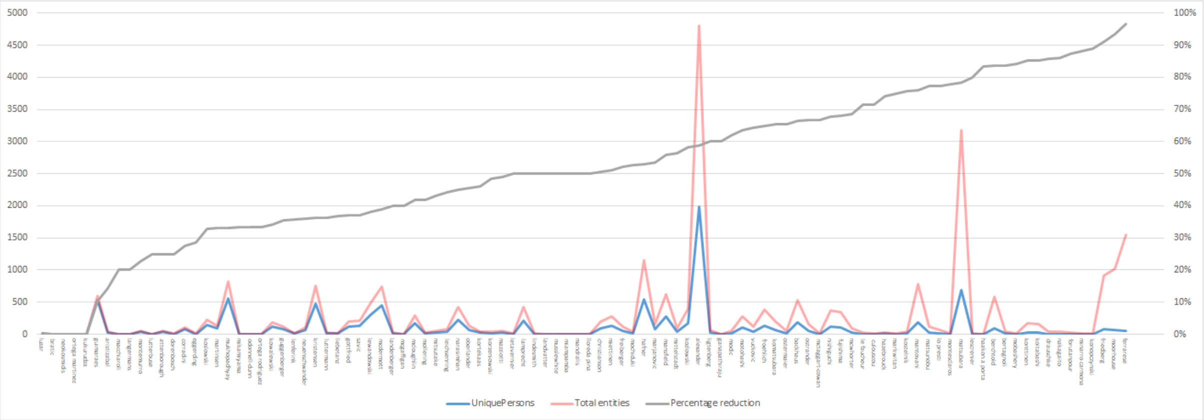


Figure 14. Reduction in unique persons

The table displays the success rate at disambiguating persons. On the left axis we have the number of distinct persons and the right hand side displays the percentage reduction when comparing the drop in the number of persons after the disambiguation with the number of persons partaing to a certain number of surnames before the disambiguation. We include on the right hand also an example of a surname, where we were not successful in advancing the disambiguation (surname Luzar), but omit the rest of such surnames in the graph. On the right hand side is a surname (surname Ferrarese) where our algorithm decreased the number of unique persons by 96 %.

Interconnected persons across EPO LOD and SN LOD - results

We have selected 100 randomized surnames and checked the level of interconnection between EPO and SN LOD persons. The rate of successfully interconnected 9,843 unique persons (i.e. persons after disambiguation), included in these 100 surnames, is approx. 1,9%[30].

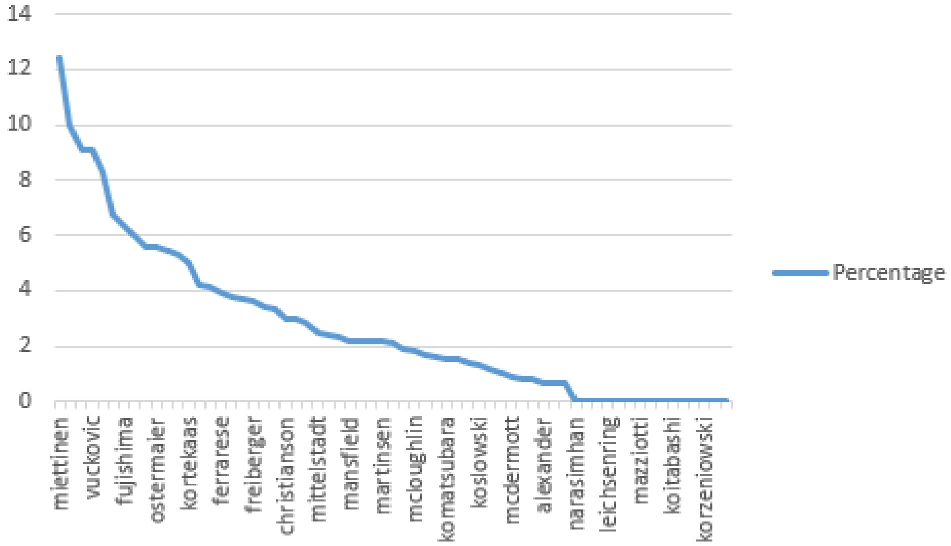


Figure 15. Percentages of interconnected persons

Percentages of interconnected persons including only surnames with above 100 unique persons.

[30] We have been conservative in this calculation, in terms of the list of these random surnames including also surnames where there is only one person available to us within some last name (to match the whole dataset), hence connecting them to another person within surname is not possible, as no additional person exists (e.g for surname Le Pochard there is only one EPO pre-disambiguation person, with no SN LOD pre-disambiguation persons).

Precision and recall

We provide a precision and recall graph (Figure 16) and the confusion matrix (Table 8).

The algorithm was run for all thresholds values from the interval [0.1, 0.9] with a 0.05 step. Figure 16 presents the evolution of the algorithm through lines on the precision-recall graph for several different blocks of data as well as a cumulative value. The prediction rate reached maximum at threshold 0.5 with 72.6% accuracy for the testing set, precision of 0.609, and recall of 0.981. Compared to the precision 0.679 and recall of 0.451 for the simple string match, the algorithm presents a significant improvement in both precision and recall.

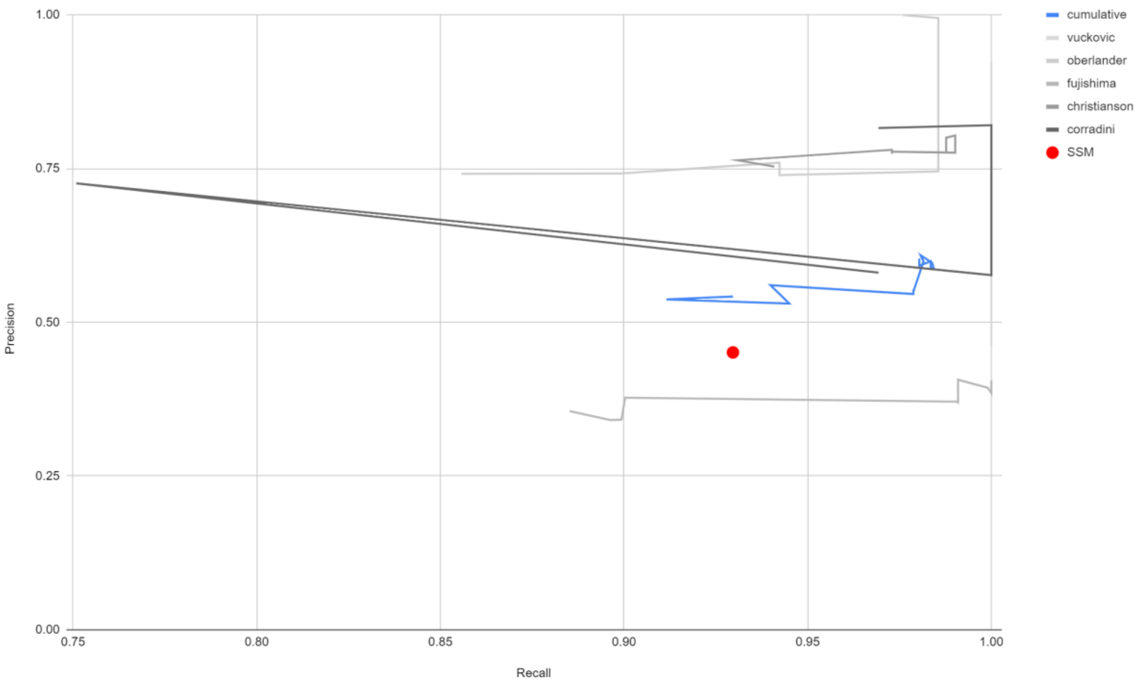


Figure 16. Precision and Recall rate for IP LodB ‘golden’ dataset.

We have been very conservative in regard to our sample, when doing the calculations; and we do not include trivial cases.[31] This means that we only include the names which are likely to be matched; either having the same last name and same first name; or having same last name and same initial, e.g., we would include for Modic, Dolores as a potential non-match (to determine the rates for false negatives) Modic, D., but would not include Modic, Klaus.

[31] Note if we would compare all names inside same last name for possible (also negative) matches, than the sample would have 115,319 records, where our rate for false negatives would be 1.93% and for false positives would be 1.98%, in comparison to SSM’s 32.11% for false negatives and 1.88% for false positives.

The Confusion Matrix is a performance measurement for the binary machine learning classification problem. It is a table with 4 different combinations of predicted and actual values.

Table 8. Confusion matrix of results for algorithm at threshold = 0.5

Actual/Prediction	Same	Different	Sum
Same	2495	49	2544
Different	1601	1886	3487
Sum	4096	1935	6031

Data publishing module

The IPLodB also provides publishing of the handled data. The publishing is an iterative process, and the database is being continuously filled up. We provide a graphical user interface at iplod.io for viewing our matches between EPO and SN data and downloading these in NTriples format.

The graphical user interface with search query

The graphical user interface of the web application provides an entry point for users to access our data. On the page Aligning->Person Records, we provide searching over all persons we identified as unique entities based on the datasets we are using to find matches. By default, we always present results of the latest stable version of our algorithm (currently, we are providing the version 110). The search can be made by: (i) the family name (when the search parameter is submitted, it is cleaned and so the search works in the same way for, e.g., “Vučković” or “vuckovic”), or (ii) by IPLodB ID (currently, we only provide search within the stable version, so IPLodB ID’s of persons matched in other versions are not found).

The search result is a list of IPLodB unique persons matching the search criterium. Each of them has a URI (by the LOD standard), on which we provide a union of data belonging to the persons that have been matched into the unique persons by our algorithm. When available, we also provide further links to the sources, from which the particular part of data has been retrieved. We are currently providing:

1) EP LOD: links to inventor vc’s (i.e. virtual cards), links to individual patent records, links to collaborating applicants’ vc cards.

2) SN LOD: links to SN person overview, links to publications (articles, books, chapters) records.[32]
3) GRID: links to affiliation records.
4) Crossref: "on the fly" search for further citations using the API opportunities they provide.

On every IPLodB unique person's page, for registered users, we also provide an annotation form, where the user can label the unique person as correctly matched, provide missing matches, or add other information that could be useful in further adaptations of our algorithm.

Providing RDF linkages

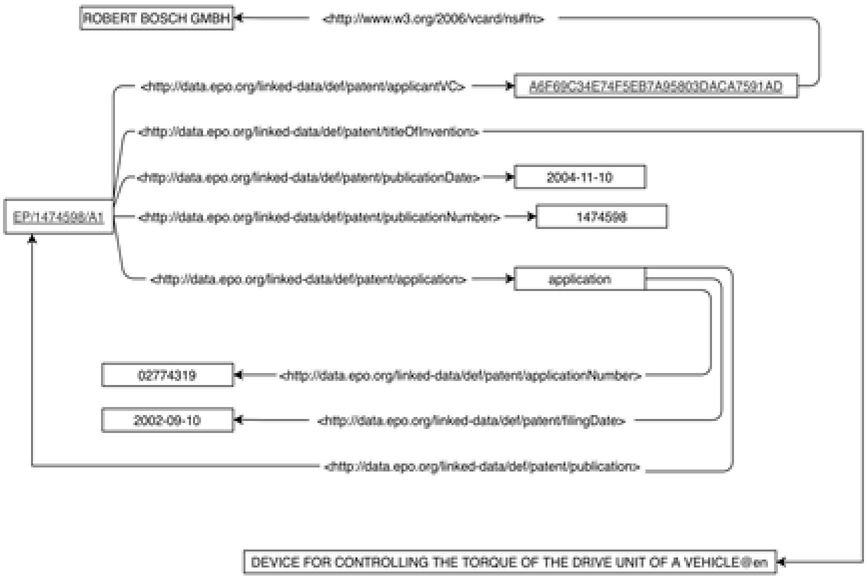

ROBERT BOSCH GMBH
<http://www.w3.org/2006/vcard/ns#fn>
<http://data.epo.org/linked-data/def/patent/applicantVC>
A6F69C34E74F5EB7A95803DACA7591AD
<http://data.epo.org/linked-data/def/patent/titleOfInvention>
<http://data.epo.org/linked-data/def/patent/publicationDate>
2004-11-10
EP/1474598/A1
<http://data.epo.org/linked-data/def/patent/publicationNumber>
1474598
<http://data.epo.org/linked-data/def/patent/application>
application
02774319
<http://data.epo.org/linked-data/def/patent/applicationNumber>
2002-09-10
<http://data.epo.org/linked-data/def/patent/filingDate>
<http://data.epo.org/linked-data/def/patent/publication>
DEVICE FOR CONTROLLING THE TORQUE OF THE DRIVE UNIT OF A VEHICLE@en

Figure 14. Interconnection inside a selected EPO LOD record of a patent

To further enhance the LOD space, we provide linkages that lead to EP LOD and/or to SN LOD. RDF (Resource Description Framework) is used for modeling information. It is an infrastructure that enables the encoding, exchange and reuse of structured data, developed by the World Wide Web Consortium (Miller, 1998; Pan, 2009). Within the framework, we however have several serialization models, e.g. N-Triples, Turtle, JSON-LD etc. In regard to our two main LOD databases, the EP LOD uses N-triples, whereas SN LOD relies on the json format. RDF is based on the idea of making statements about resources in the format of *subject - predicate – object* (see also Figure 14).

[32] Note that if grant data is provided, there are also hyperlinks leading to records related to grants.

Table 9. Exemplary links between EPO, SN and IPLodB

Triple for EPO - IPLOD
<https://data.epo.org/linked-data/data/vc/D5890E976016F4B6990707BD1D049EB5> <**http://www.w3.org/2002/07/owl#sameAs**> <https://www.iplod.io/Persons/UniquePerson/3d546f57-7103-40ab-8b6e-012978829e89>
Triple for SN - IPLOD
<https://scigraph.springernature.com/person.0664161767.02> <**http://www.w3.org/2002/07/owl#sameAs**> <https://www.iplod.io/Persons/UniquePerson/3d546f57-7103-40ab-8b6e-012978829e89>
Compilation of triples
<https://data.epo.org/linked-data/data/vc/D5890E976016F4B6990707BD1D049EB5> <**http://www.w3.org/2002/07/owl#sameAs**> <https://www.iplod.io/Persons/UniquePerson/3d546f57-7103-40ab-8b6e-012978829e89> <https://scigraph.springernature.com/person.0664161767.02> <**http://www.w3.org/2002/07/owl#sameAs**> <https://www.iplod.io/Persons/UniquePerson/3d546f57-7103-40ab-8b6e-012978829e89>

The first triple represents a link (i.e., using the standard predicate for records denoting same world objects: “Owl#sameAs”) between an EPO record and an IPLodB record, and the second for the link between an SN record and an IPLodB record. The third shows that for interlinked persons, there is also a path between EPO and SN records.

In essence RDF is an effort to make the world wide web not only “machine-readable”, but also “machine-understandable” (Lassila, 1998), which makes the standardization also the cornerstone of the linked open data. The Web Ontology Language OWL (W3C, 2020) defines the predicate owl:sameAs used for expressing equivalences between seemingly different individuals. We used the dotNetRDFs’ (2020) library to construct RDF data by linking EP and Springer LOD persons to our internal joined person references using the owl:sameAs predicate. The data is provided as a bulk download in NTriples format. We plan to provide all data also in a downloadable form of bulk downloads with provided triplets between EP LOD, SN LOD and the IPLodB as soon as the IPLodB related publication efforts are completed. We will use the owl#sameAs predicate to connect the records in these databases (see Table 9); and would encourage both providers to also use these links inside their dataset. Sample data is available on iplod.io and in the meantime the users can also query the data via search query on iplod.io.

Discussion

The overall objective of the IP LodB project has been to explore and develop connections from EPO LOD bibliographic data to other LOD sources available on the world wide web. The IP LodB project approached the objective in two ways: 1) from the top-down, the IPLodB Map is a qualitative evaluation of LOD sources (presented under separate cover) and 2) from the bottom up, The IP LodB Database is a quantitative evaluation of linkability between EPO LOD and SN LOD (presented in this document). Previous research using bibliographic data suggests that, while lists of publications are unique, the lists of people who wrote them are not and may contain incomplete, and sometimes inconsistent, text strings. Our descriptive exploratory analyses and plots of the data from both LOD sources suggests that both:

1) Entity resolution for people is weak and
2) Non-ignorable missing and ambiguous data exist.

We have described in utter detail the the IP LodB Database intervention to both:

1) Validate and augment the LOD data from other sources and
2) Increase the uniqueness of both LOD sources by introducing a *nascent* person identifier based on disambiguation procedures for name text strings, i.e., for inventors and authors, and other person attributes, e.g., collaboration, affiliation, location, and content.

Our approach has contributed both conceptual clarity and technical innovation to the name disambiguation literature. On the conceptual side, we have reviewed the scientific concepts and operational definitions used in previous research on name disambiguation for bibliographic data. Based on our review of the literature, we propose a relatively parsimonious model for disambiguating people using their family and given names, and four generic categories of other attributes, e.g., collaboration, affiliation, location, and content. We have argued why we expect these constructs to be convergent and discriminant valid, at least in principle. To some extent, our theory is validated by examples revealed in the Results section, where an indicator for each category has been shown to be sufficient to disambiguate an ambiguous name text string. However, there remain some counterexamples where our procedure breaks down and produces false positives. Nevertheless, we have steadily reduced lumping errors during the project and are confident of reducing them further as we refine our procedure.

On the technical side, we have proposed a creative application of conventional RDBMS technology to add value to leading edge LOD technology, which is pivotal to democratize the availability of data. Specifically, to detect matches between person-publication records that refer to the same person with high probability, and distinguish person publication records with similar names, but that refer to different people, for the large number of records we are dealing with in our project, several steps needed to be followed. First, the data should be stored efficiently, i.e., all attributes for a single entry should be retrieved fast, and due to dynamic content (some attributes may be updated), they should be stored without repeating (as is the case in some document-oriented NoSQL databases). For these reasons, we store the data in a relational SQL database. Second, there is the problem with missing data; not all records contain values of all attributes, but in many cases the union of data for three or four records contains a lot of data on other attributes that can help join the records with only partial data. For this reason, we use an iterative matching approach. Third, due to many records, we use a standard blocking technique, which enables fast matching of records and, thus, reduces the number of single entities to a more manageable size. Finally, we release the data back to the wide world web in two formats that bridge the datasets; one is through a user interface and with search query, including a wiki-type dialogue box for users to propose edits, where the links lead back to LOD source data (not only EPO LOD and SN LOD, but also GRID) and the other in the format of RDF links in line with our data publishing module, for already processed blocks.

IP LodB has also provided an entry to EP LOD, which is a free and open database, for:

1) More advanced (experienced with IP and LOD) users to have a user friendly entry point to existing LOD data, already connected to additional LOD, which they can use to understand the scope of the content and evaluate its usability without necessarily engaging in download of the dataset per se.
2) Users without prior LOD knowledge can gain an user friendly insight into LOD data, which is otherwise optimized for machine-readability.

However, the project provides benefits beyond this. For example there is also an opportunity to delve into particular fields, especially emerging fields that are high on public policy agendas, e.g. Circular Economy (CE). CE is likely to be built on basic research effort and applied innovation. Japan and China lead in intellectual property registrations for CE related innovations with South Korea and Europe active, but trailing in CE patent stock. Exploring new opportunities in developing innovation metrics, thus provides an interesting aspect, for a concept that has been called both fuzzy and an umbrella concept; including various activities (from re-cycling, to re-using, re-furbishing, etc.), which are hard to capture with conventional innovation metrics. Three most usual search approaches to identify patents in a particular field are those of using two sets of search terms: domain keywords, patent classifications (IPC or CPC codes) and, if applicable, the so-called Y-codes. We believe our efforts inside IP LodB allow us to further adapt the RAKE algorithm and allow us to observe the flow of inventive ideas across publication and patenting activities by using a consolidated patent-publication database.

Furthermore, the IPLOD data allows, similar to other patent data, to capture changes in time, e.g. to track person’s migration between organizations and the evolution of their field of work (see Figure 15) via clickable links leading all the way to original sources (including the EP patent documents).

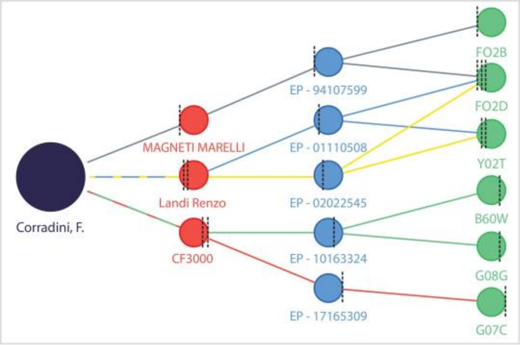


Figure 15. Inventive activity through different attributes and time

The blue circle represents the applicants he collaborated with, the blue dots are the patents, and the green dots the CPC codes. The radius of the circle represents the span of approximately 20 years this inventor has been active. The dotted line represents the time stamps, showing for example that his focus of work migrated from the mid-90s from F02 (working on computations for internal combustion engines systems) to G07 (developing remote testing systems for registering vehicle performance). For this person the han_id (which is one of the disambiguation efforts for patent data included in the Patstat) for example do not match. The psn_id (which marks another disambiguation effort included in Patstat) do match, which is consistent with the approach of exact same name matching.

The additional benefit of using IPLOD data being our additional disambiguation efforts, which can be useful also to those interested not only in EPO LOD, but also Patstat. We add to the current two disambiguation efforts that are included in the Patstat (Global) database, and focus predominantly on organizations (applicants). Note that the database includes the PSN_ID and HAN_ID also for all "Persons", defined in Patstat as legal persons or natural persons[33], albeit both other disambiguations are mainly focused on applicants. Nonetheless, many natural persons in Patstat do have a PSN_ID, with indication that these names have indeed been subjected to some degree to the related harmonization process[34]. However, these disambiguation procedures were not optimized for natural persons (including those that are applicants) inside Patstat. We believe our disambiguation could provide an additional disambiguation element of higher quality in terms of natural persons, either as applicants or inventors to Patstat as such, or to Patstat users. In the future we believe we could also be able to provide additional value for the disambiguation of organizations.

Furthermore, IP LodB data can also show how the inventors build diverse collaboration networks in their publishing and inventorship activities; as well as their connections with distinct organizations (i.e. affiliations). Consider the following case. The person has been disambiguated from several SPR and EPO persons. The SPR only has the surname comma initial for his name, whereas for EPO records both first and last names are available in full. We can observe his wider publication collaborations with 50 authors altogether, but only collaborates on patents with 6 inventors (see Figure 16). For all but one publication, his affiliation (or at least one of them) is a higher education institution, only one

[33] With the distinction made by using the fields APPLT_SEQ_NR > 0 (indicating an applicant) and INVT_SEQ_NR > 0 (indicating an inventor. Of course natural persons, which are our focus, can appear as both inventors and applicants.

[34] The PSN_ID is for PSN_NAMEs which have been created during the harmonisation process in the range 1...100 000 000 (Patstat, 2019).

is connected to a company. All the applicants are however the company, and his inventorship activities remain unconnected to the higher education institution.

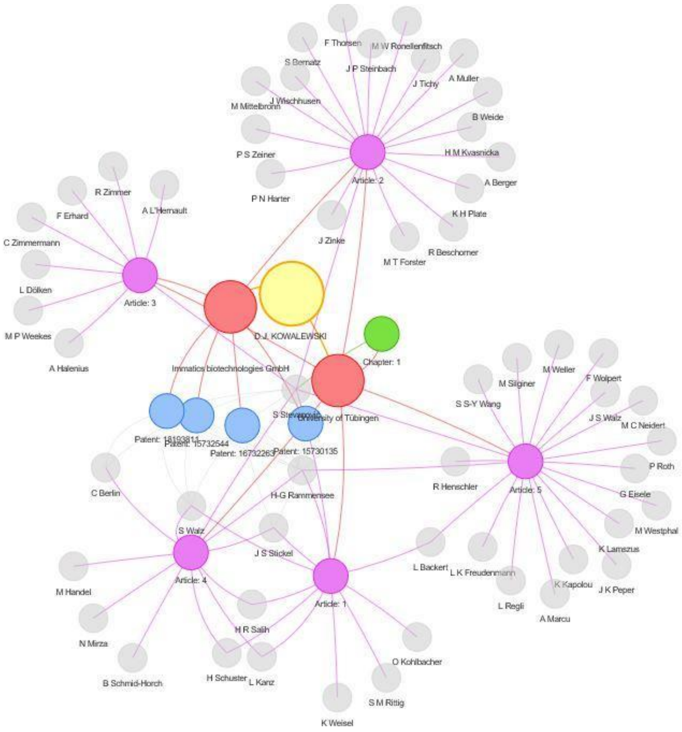


Figure 16. Publishing and inventive activity

The Figure depicts an inventor's (yellow circle) publishing (articles violet, chapters green circles) and patenting (blue circles) activity (note: with blue circles including all patents with unique URI). We can see that he also has two distinct affiliations (red circles) and a myriad of coauthors (grey circles). There is a dense central part, where the patenting activity and some of his collaborating coauthors/coinventors are.

Another benefit of the linked open data and our bridging approach comes to the forefront here. For the patent record EP- 15730135 the original provider of the EPO LOD patent data, has provided additional information, in the form of publication data, where we can see that the patent has been granted (so-called B1 document or the European patent specification, i.e., granted patent) recently, i.e. on April 29th 2020. Thus we can see that by bridging the records, this allows the original providers of the data to add further details to a record; and the given link between EPO LOD and SN LOD will allow for discovery of new data, since URIs are stable.

Limitations and future work

We still need to apply our method to all blocks in the IP LOD database. The time for processing one block depends on its size, however, even compiling the data and iterative comparisons of similarities for all persons within a relatively small block is taking some considerable amount of time (up to several minutes), and there are about a million of such blocks. However, blocking enables high parallelization of the processes, and thus many blocks can be evaluated at the same time. There is also an issue of handling extremely large blocks, where the iterative process is taking much more time. In such cases some additional "internal blocks" could be introduced. As mentioned earlier, one of the future tasks remains to implement iterative blocking also as a wrapper of the existing methods, meaning that after the first complete evaluation of the database, new blocking (using other attributes as filter parameters) takes place and tries to find potential matches.

There are, however, still challenges to be addressed in our method. Its iterative nature on one side enables collecting a lot of data about one entity, which has a negative side once a false positive is introduced, e.g., when we match two entities that should not be matched. In such a case, we collect the data of two distinct persons and we compare all other entities against such a (wrong) package of data, resulting in more false matches. However, due to our tight and conservative approach, such cases appear rarely, and further enhancements of the method should reduce them even more. However, introducing false positives is perhaps a bigger problem than introducing false negatives, i.e., not matching entities that should be matched, since in this way we do not introduce new false facts, while having not matched pairs only means there is a lack of data about the entities and they might be matched when additional information will be obtained.

Our selection of both main datasets (EPO LOD and SN LOD) is not without limitations. SN LOD only includes a subset of all publication activity by individuals (those connected to Springer Nature publishings), the EPO LOD on the other hand in their non-EPO records only provide limited information (and sometimes does not contain information on inventors). However, our work on LOD MAP can direct the efforts towards other databases that could be interesting to add; with our immediate focus on the *dblp computer science bibliography* (DBPL) database[35].

Furthermore, also in terms of the included variables, more could be done. For example, EPO does not include in the EPO LOD dataset the Person identification (person_id), which would provide an easier way for users of the LOD to connect it to the Patstat record, should they wish to derive more information from Patstat (e.g. on Person address). Furthermore, there would be a benefit to providing the patent documents' abstracts also in the LOD version. This would allow the users to have the best of both worlds; the data richness of the Patstat, as EPOs main product, and the linked open data approach. Doing so, would be an additional way to enhance the usability of the EPO LOD dataset.

Matching persons is of course just the first step of making a useful IP LOD subcloud with EPO LOD as a hub, which uses a diligent method of aligning datasets. Using the same methodology, matching can be expanded to companies/organizations, providing an efficient tool of browsing through various attributes, patents being only a small portion of provided information. Evidently, there is a need for such a service, due to sparql limitations (see 2.1 deliverable for more) and new data sources from reputable publishers (also expected datasources, such as the new planned LOD catalogue from Eurostat). In the long run, the IPLOD could also be a common entry point for EPO related research results, especially for those from EPO Academic Research Programme (EPO ARP) research projects, which are now scattered across different repositories, and are more or less visible.

[35] Openly available at https://dblp.org/

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer: New York, NY.

Bryl, V., Bizer, C., Isele, R., Verlic, M., Hong, S. G., Jang, S., Yi, M. Y., & Choi, K.-S. (2014). Interlinking and knowledge fusion. In S. Auer, V. Bryl, & S. Tramp (Eds.), *Linked Open Data—Creating Knowledge Out of Interlinked Data: Results of the LOD2 Project* (pp. 70–89). Springer.

Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge: New York, NY.

EPO. (2019). *Linked Open Data from the European Patent Office*. European Patent Office (EPO) - Linked Open EP Data. epo.org/linked-data

EPO. (2020). *PATSTAT - Backbone Data Set for Statistical Analysis*. European Patent Office (EPO) - PATSTAT. http://www.epo.org/patstat

GeoNames. (2020). *GeoNames—Geographical database covers all countries*. GeoNames. https://www.geonames.org/

GRID. (2020). GRID STATISTICS. Available at: https://www.grid.ac/stats

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer: New York, NY.

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1), 414-427.

Idrissou, O. A. K., van Harmelen, F., & van den Besselaar, P. (2020). Network Metrics for Assessing the Quality of Entity Resolution Between Multiple Datasets. *Semantic Web, in press*. https://doi.org/1570-0844/0-1900

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical*

Learning: With Applications in R. Springer: New York, NY.

Khusro, S., Jabeen, F., Mashwani, S. R., & Alam, I. (2014). Linked Open Data: Towards the Realization of Semantic Web- A Review. *Indian Journal of Science and Technology*, *7*, 20.

Lassila, O., & Swick, R. R. (1998). Resource description framework (RDF) model and syntax specification.

Li, G.-C., Lai, R., D’Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., Yu, A. Z., & Fleming, L. (2014). Disambiguation and co-authorship networks of the U.S. patent inventor database (1975–2010). *Research Policy*, *43*(6), 941–955. https://doi.org/10.1016/j.respol.2014.01.012

Magerman, T., Looy, B. van, & Song, X. (2006). *Data production methods for harmonised patent statistics: Patentee name harmonisation* (pp. 1–92). Eurostat Working Paper and Studies: Luxembourg.

Maraut, S., Dernis, H., Webb, C., Spiezia, V., & Guellec, D. (2008). *The OECD REGPAT Database: A Presentation* (Science, Technology and Industry Working Papers No. 02; pp. 1–37). OECD: Paris, France. https://www.oecd-ilibrary.org/science-and-technology/the-oecd-regpat-database_241437144144

Martínez, Gema Lax, Julio Raffo, and Kaori Saito. (2016). Identifying the Gender of PCT Inventors. WIPO Economic Research Working Paper 33. World Intellectual Property Organization - Economics and Statistics Division. Available at: http://www.wipo.int/publications/en/details.jsp?id=4125.

Miller, E. (1998). An introduction to the resource description framework. Bulletin of the American Society for Information Science and Technology, 25(1), 15-19.

Morrison, G., Riccaboni, M., & Pammolli, F. (2017). Disambiguation of patent inventors and assignees using high-resolution geolocation data. *Scientific Data*, *4*(1), 170064.

https://doi.org/10.1038/sdata.2017.64

OECD. (2019, 2002). The OECD REGPAT Database. Available at: https://www.oecd-ilibrary.org/science-and-technology/the-oecd-regpat-database_241437144144

Pan, J. Z. (2009). Resource description framework. In Handbook on ontologies (pp. 71-90). Springer, Berlin, Heidelberg.

Patstat. (2019). Patstat Global (autumn 2019 edition).

Pezzoni, M., Lissoni, F., & Tarasconi, G. (2014). How to kill inventors: Testing the Massacrator© algorithm for inventor disambiguation. *Scientometrics*, *101*(1), 477–504. https://doi.org/10.1007/s11192-014-1375-7

Raffo, J., & Lhuillery, S. (2009). How to play the “Names Game”: Patent retrieval comparing different heuristics. *Research Policy*, *38*(10), 1617–1627. https://doi.org/10.1016/j.respol.2009.08.001

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic Keyword Extraction from Individual Documents. In M. W. Berry & J. Kogan (Eds.), *Text Mining: Applications and Theory* (pp. 3–20). Wiley: Chichester, UK. https://doi.org/10.1002/9780470689646.ch1

Sleeman, J., Finin, T., & Joshi, A. (2015). Entity Type Recognition for Heterogeneous Semantic Graphs. *AI Magazine*, *36*(1), 75–86. https://doi.org/10.1609/aimag.v36i1.2569

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology*, *43*(1), 1–43. https://doi.org/10.1002/aris.2009.1440430113

SN. (2019). *SN SciGraph: A Linked Open Data Platform for the Scholarly Domain*. Springer Nature (SN) - SciGraph. https://www.springernature.com/scigraph

Torvik, V. I., & Smalheiser, N. R. (2009). Author Name Disambiguation in MEDLINE. *ACM*

Trans. Knowl. Discov. Data, *3*(3), 11:1–11:29. https://doi.org/10.1145/1552303.1552304

W3C. (2020). 5.2.1 owl:sameAs. Available at: https://www.w3.org/TR/owl-ref/#sameAs-def.

Wenzel, R., & Van Quaquebeke, N. (2018). The Double-Edged Sword of Big Data in Organizational and Management Research: A Review of Opportunities and Risks. *Organizational Research Methods*, *21*(3), 548–591. https://doi.org/10.1177/1094428117718627