



European Spatial Data Research

July 2019

EuroSDR - EuroGeographics Seminar
Data Linking by Indirect Spatial
Referencing Systems

September 5th - 6th 2018 - Paris, France

Bénédicte Bucher, Esa Tiainen, Thomas Ellett,
Elise Acheson, Dominique Laurent, Sylvain Boissel

Workshop Report

The present publication is the exclusive property of
European Spatial Data Research

All rights of translation and reproduction are reserved on behalf of EuroSDR.
Published by EuroSDR

EUROPEAN SPATIAL DATA RESEARCH

PRESIDENT 2018 – 2020:

Paul Kane, Ireland

VICE-PRESIDENT 2017 – 2019:

Fabio Remondino, Italy

SECRETARY – GENERAL 2019 – 2023:

Joep Crompvoets, Belgium

DELEGATES BY MEMBER COUNTRY:

Austria: Michael Franzen, Norbert Pfeifer

Belgium: Eric Bayers

Croatia: Ivan Landek, Željko Bačić

Cyprus: Andreas Sokratous, Georgia Papathoma, Andreas Hadjiraftis, Dimitrios Skarlatos

Denmark: Jesper Weng Haar, Tessa Anderson

Estonia: Tambet Tiits, Artu Ellmann

Finland: Juha Hyypä, Jurkka Tuokko

France: Bénédicte Bucher, Yannick Boucher

Germany: Paul Becker, Lars Bernard

Ireland: Paul Kane, Audrey Martin

Norway: Jon Arne Trollvik, Ivar Maalen-Johansen

Poland: Piotr Woźniak, Krzysztof Bakula

Slovenia: Dalibor Radovan, Peter Prešeren, Marjan Čeh

Spain: Julián Delgado Hernández, Javier Barrado Gozalo,

Sweden: Tobias Lindholm, Thomas Lithén, Heather Reese

Switzerland: André Streilein, François Golay

The Netherlands: Jantien Stoter, Martijn Rijdsdijk

United Kingdom: Sally Cooper, Claire Ellul

ASSOCIATE MEMBERS AND THEIR REPRESENTATIVES:

Esri: Nick Land

Informatie Vlaanderen: Jo Van Valckenborgh

nFrames: Konrad Wenzel

Terratec: Leif Erik Blankenberg

Vexcel: Michael Gruber

ISpatial: Dan Warner

COMMISSION CHAIRPERSONS:

Data Acquisition: Jon Mills, United Kingdom
Modelling and Processing: Norbert Haala, Germany
Updating and Integration: Jon Arne Trollvik, Norway
Information Usage: Bénédicte Bucher, France
Business Models and Operation: Joep Crompvoets, Belgium
Knowledge Transfer: Markéta Potůčková, Czech Republic

OFFICE OF PUBLICATIONS:

Bundesamt für Eich- und Vermessungswesen
Publications Officer: Michael Franzen
Schiffamtsgasse 1-3
1020 Wien
Austria
Tel.: + 43 1 21110 825200
Fax: + 43 1 21110 82995202

CONTACT DETAILS:

Web: www.eurosdrr.net
President: president@eurosdrr.net
Secretary-General: secretary@eurosdrr.net
Secretariat: admin@eurosdrr.net

EuroSDR Secretariat
KU Leuven Public Governance Institute
Faculty of Social Sciences
Parkstraat 45 bus 3609
3000 Leuven
Belgium
Tel.: +32 16 37 98 10

The official publications of EuroSDR are peer-reviewed.

Bénédicte Bucher, Esa Tiainen, Thomas Ellett, Elise Acheson,
Dominique Laurent, Sylvain Boissel

EuroSDR – EuroGeographics Seminar Report

DATA LINKING BY INDIRECT SPATIAL REFERENCING SYSTEMS

September 5th - 6th 2018 – Paris, France

Index of Figures	6
1 INTRODUCTION	7
2 PRESENTATIONS SUMMARY	8
2.1 Gazetteers for linking text to space: experiences with contrasting corpora, Elise Acheson, University of Zurich.....	8
2.2 Georef - Service and Development platform: Research data pilot overview, Esa Tiainen, National Land Survey of Finland.....	10
2.3 Assessing the importance of named places: benefits and difficulties, Dominique Laurent, IGN France	11
2.4 Designing Data projects, how to value geographical heritage data with state of the art solutions?, Julien Homo, Kévin Darty, Foxcub	12
2.5 Finnish Linked Data pilots, Kai Koistinen, National Land Survey of Finland.....	13
2.6 The challenge of linking or integrating data on Buildings, Dominique Laurent, IGN France	15
2.7 Administrative Units as Linked Open Data – A case study from the Norwegian Mapping Authority, Thomas Ellett, Kartverket.....	16
2.8 Wikidata, a short introduction, Julien Boissel, Wikimedia foundation	17
2.9 Linear indirect reference systems to interconnect data in transportation applications, Alain Chaumet, ENSG-Valilab.....	18
3 COMMON FINDINGS ISSUED FROM WRAP UP SESSIONS	20
3.1 Need for ontologies of places and of digital assets	20
3.2 Consistency in an open world.....	21
3.3 Computing, maintaining and sharing links	21
3.4 Communities, commitments, authorities	22
4 CONCLUSION AND FURTHER AREAS OF RESEARCH AND DEVELOPMENTS	23
5 ACKNOWLEDGEMENTS	24
ANNEX 1 – PROGRAM	25

EuroSDR – EuroGeographics Seminar Report

DATA LINKING BY
INDIRECT SPATIAL REFERENCING SYSTEMS

September 5th - 6th 2018 – Paris, France

With 17 figures

Bénédicte Bucher^{a, b}, Esa Tiainen^c, Thomas Ellett^d, Elise Acheson^e,
Dominique Laurent^a, Sylvain Boissel^f

^a IGN-France

^b University Paris Est, LaSTIG

^c Maanmittauslaitos, Finland

^d Kartverket, Norway

^e University of Zurich, Switzerland

^f Wikimédia France

Index of Figures

Figure 1: Linking text to space is a 3-step processing pipelines, where each step can use gazetteers (© Acheson).....	8
Figure 2: Application of a Named Entity Recognition tool, © Acheson, from https://www.nytimes.com/2018/07/11/science/hominins-tools-china.html	9
Figure 3: In the Finnish Georef pilot application, place names are used to bridge information (research reports, pictures) and data assets (geological surveys). © Tiainen	10
Figure 4: Georef architecture © Tiainen.....	10
Figure 5: Designing a map requires selecting place names and deciding how to portray them depending on their importance but also on the cartographic context which includes the density of information and the length of the name. (sources: IGN Top100 and Top25 maps).....	11
Figure 6: INSPIRE conceptual model for NamedPlace. A named place has several attributes relating to its importance. However, these are hardly reliable because subject to interpretation and in practice often void.	11
Figure 7: UN GGIM core data model for named place. A named place should be captured with its true geometry and its population (for populated places); in addition, the geometry reliability should be documented. © Laurent. Sources of the maps: IGN Top100, Top 25, web.	12
Figure 8: Snapshots from Navigae platform developed by Foxcub to locate geographical documents, to interconnect them with Web content like Wikipedia (a) and to support visual inter-comparison (b).	13
Figure 9: Foxcub matrix to select the most relevant technologies in the course of the project © Foxcub.....	13
Figure 10: A http data card from the Finnish pilot © Kostinen.....	14
Figure 11: Geographical names are a glue between different datasets in Finland. SU: statistical unit, AU: administrative unit, GN: geographical name, BU: building, AD: addresses © Kostinen	14
Figure 12: Categories of information related to Buildings in the INSPIRE data specifications that might be considered in a unifying framework. © Laurent	15
Figure 13: different views (segmentations) on buildings, even in a simple case © Laurent	16
Figure 14: Pros (5 to 8) and cons (1 to 4) of Linked Data, lessons learned after Kartverkets pilot. © Ellett	17
Figure 15: Projects belonging to the Wikimedia movement comprise content project (Wikipedia, Wiktionary, Wikibooks, Wikiquote, Wikivoyage, Wikisource, Wikimedia Commons, Wikispecies, Wikinews, Wikiversity, Wikidata) and infrastructure and coordination projects (Meta-Wiki, Wikimedia Incubator). © Wikimedia.....	17
Figure 16: The item Douglas Adams in Wikidata. Source: Wikimedia, © CCO.....	18
Figure 17: Reversible transformation from linear location to geographic coordinates spatial location (direct location) © Chaumet	19

1 INTRODUCTION

An indirect spatial reference (ISR) is any way to describe a location without using coordinates. It can be the name of a located feature (for example "Zürich") or a code that identifies a located feature (such as a zip code). Place names in particular are one of the most ancient spatial referencing frameworks, to reference a person by the name of her birth place, to reference a legal document by the name of the place it was signed. As opposed to coordinates, indirect spatial referencing is well adapted to provide references readable by humans. The scientific and technological evolutions in information infrastructures come with new opportunities for using indirect spatial referencing systems to interconnect pieces of information. Indeed, every feature that has an unambiguous identifier and an unambiguous location (e.g. a named place, an address, an administrative unit, a building) can be used for the indirect spatial referencing of other features, and more and more features satisfy such conditions; they have unambiguous identifiers because of the semantic web and they have unambiguous location thanks to geocoding services and to collaborative platforms where people can discuss and reach some consensus. The evolutions in information infrastructures also bring new requirements. The growing corpus of textual content produced by society, where spatial information is an important dimension for analysis, has led to more and more communities using indirect spatial referencing systems with new specific requirements stemming from their work.

In this context, the seminar reported here aimed at presenting and discussing results and pending issues from the literature as well as practitioners' experience in the domain of using indirect spatial referencing systems or designing them. It was co-organised by EuroSDR Commission 4 on information usage and by EuroGeographics Knowledge Exchange Network INSPIRE, in September 2018. It gathered 17 participants from 7 countries. The first day of the seminar covered key results and issues on data linking and referencing with place names. A keynote presented state of the art text-to-space methods, remaining challenges and requirements in terms of indirect spatial referencing systems. It was followed by presentations from projects and organisations in charge of designing and distributing place names databases. Some presentations targeted national scope with a priority on the usability of authoritative frameworks. Others targeted the European or international scope where national gazetteers are highly heterogeneous, with a concern for reusability and interoperability. A shared statement was the importance of improving our model of places, and of relations to places, as well as explicitly model scopes and contexts, to improve gazetteer services and handle heterogeneities. Another statement was that the competences needed to design operational applications thanks to the interconnection of a wide range of information assets -scientific papers, maps, collaborative web content, today embrace the ability to discover and use APIs, as well as the capacity to experiment and follow up technologies (e.g. triple store solutions) in order to select them. Second day enlarged the scope to other core data, like buildings, including linear referencing systems and was also devoted to wrap up and draft conclusions. A common view was that usages of the abundant assets of digital data available are still partial and restricted and data linking is a substantial way forward as it enables inter and intra domain connections through a common spatial framework. With respect to using new technologies, the presentation of different projects, including pilots, in different countries evidenced notable progression and the need for national mapping agencies to embrace new competences. Yet, this requires some investment and these investments can be difficult to fulfill when the expected benefits are theoretical

and when substantial investments have already been made over many years in other technologies.

2 PRESENTATIONS SUMMARY

2.1 *Gazetteers for linking text to space: experiences with contrasting corpora,* *Elise Acheson, University of Zurich*

Linking text to space is needed to enable spatial analysis or rendering of a text or a text corpus: visualizing news articles, searching for geographically relevant documents, reconstructing an itinerary out of sparse textual clues, modelling user location and context. Gazetteers play a central role in many text-to-space workflows, for identifying possible toponyms within texts and resolving these toponyms to a unique identifier and potentially linking to spatial representations.

In her talk, Elise Acheson from the University of Zurich, discussed requirements for gazetteers used in linking text to space. Three studies were presented to illustrate the presentation. A first study was the analysis of Swiss hiking blogs to evaluate how people perceive a set of landscapes in Switzerland, a geographically-focused corpus containing many fine-grained toponyms. A second study was the analysis of scientific articles to enable spatial hypotheses on domains covered by these papers. This corpus presented a more global, yet more common, set of locations. A third study was the analysis of twitter to geolocate a disaster or look at urban planning issues. Twitter corpora presented global locations, varying granularities, and very limited context.

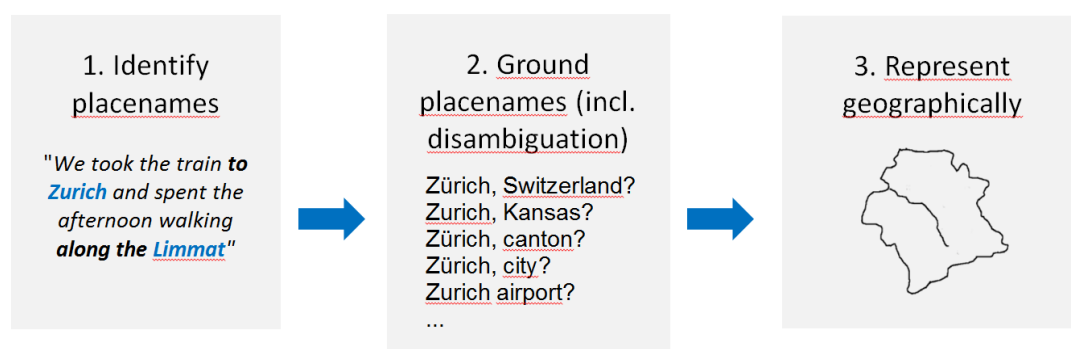


Figure 1: Linking text to space is a 3-step processing pipelines, where each step can use gazetteers (© Acheson)

Linking text to space implies three steps as summarized figure 1. Firstly, the identification of toponym in the text is covered by a combination of manual identification, gazetteer lookup, rule-based analysis and machine learning named entity recognition (Leidner and Liberman 2011). The process needs to be tuned depending on the nature of the corpus (type of place names, existence of training data, on the existence of a structure for the text). Out of the box named entity recognition (NER) tools are of high performance and can be improved if there exists training data. Figure 2 shows an output of a NER tool. Manual identification is especially useful when knowledge from local expert is necessary, like in the Swiss hiking blog study. While some texts use well identified references to place names (in..) for location, location can sometimes be described through complex expressions, like in the articles study e.g. "at a colony site near the Orbetello lagoon (Grosseto, Tuscany, central Italy)". The

Twitter study is a very complex case where the granularity of toponym varies and where colloquial place names are often used.

Input text

Chipped rocks found in western China indicate that human ancestors ventured from Africa earlier than previously believed. "The implications of all this are large," said Michael Petraglia, a paleoanthropologist at the Max Planck Institute for the Science of Human History, who was not involved in the new study. "We must re-evaluate our understanding of human prehistory in Eurasia."

Output NER-tagged text

Chipped rocks found in western **China** indicate that human ancestors ventured from **Africa** earlier than previously believed. "The implications of all this are large," said **Michael Petraglia**, a paleoanthropologist at the **Max Planck Institute for the Science of Human History**, who was not involved in the new study. "We must re-evaluate our understanding of human prehistory in **Eurasia**."

Figure 2: Application of a Named Entity Recognition tool, © Acheson, from <https://www.nytimes.com/2018/07/11/science/hominins-tools-china.html>

The next step, toponym disambiguation, is a step where there is most need for improvement. Toponymy is very ambiguous for computers and the way people think and reason about space is still difficult to reproduce in a computer. Typically, bounding of cognitive regions is imprecise and variable. In 'naïve' geography, qualitative spatial relations are not always consistent with metric measures (Egenhofer and Mark 1995). Besides, there are lots of different referring expressions, i.e. ways people write and talk about locations, and most of them can be decoded only by using the context of the expression. At the end of this second step, the text is annotated to associate to the identified toponyms an annotation specifying unambiguously the spatial entity referenced by this toponym.

Last step is the design of a spatial footprint for the document or for the corpus based on step 2 and the geometries associated to toponyms, retrieved thanks to a geocoder and then processed (aggregated for instance).

An important limitation of the domain underlined by Elise Acheson is the representation of spatial features and properties of these references, as well as spatial relationships between them. Whereas these categories of information classically are worn by geometry, named places usually are represented as points; with no scale information, no representation of vagueness, not always a connection to spatial objects nor possibility to evaluate spatial relationships between places. Solutions may involve multiple-representation, explicit hierarchical information for simple spatial reasoning. Another limitation is the lack of sufficient information about the scope of a gazetteer and the wide variety of gazetteer coverage and quality. Some important information often is implicit or unknown: when is a name official, to which feature types apply which rules, etc. Solutions could involve focusing on data provenance metadata delivery, and more information for non-expert users. The variety of contexts of text design and corpus usage also is an issue because they impact the relative importance of gazetteer records. A solution could be to make more use of data (query logs, social media) to propose a measure of importance that adapt to the user context.

References

Egenhofer M.J., Mark D.M. (1995) Naive Geography. In: Frank A.U., Kuhn W. (eds) Spatial Information Theory A Theoretical Basis for GIS. COSIT 1995. Lecture Notes in Computer Science, vol 988. Springer, Berlin, Heidelberg

Leidner, J. L., & Lieberman, M. D. (2011). Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. SIGSPATIAL Special, 3(2), 5.11.

2.2 Georef - Service and Development platform: Research data pilot overview, Esa Tiainen, National Land Survey of Finland

Georef is an initiative for a service and application development platform that employs http URIs of place names for geocoding different data assets to enable and improve data combinations of spatial data and any other data using linked data technology. Place names can be used for bridging different information and data assets as a lot of information carry place names but most information does not carry direct location data.

Since a first pilot in 2015, the Finnish National Land Survey has developed a framework for data linking by Place names. Current pilot aims at linking scientific research reports: geological deposits in research reports are annotated with geology ontology (URIs) with their location presented with place names (see figure 3). Annotation results are added to RDF database, using open-source MAUI annotation tool. A smart search tool has been developed for indexing search results from RDF with ElasticSearch (see figure 4). The user interface shows the geology deposits and those selected with keywords on map. A taxonomy for classification of results is still needed.



Figure 3: In the Finnish Georef pilot application, place names are used to bridge information (research reports, pictures) and data assets (geological surveys). © Tiainen

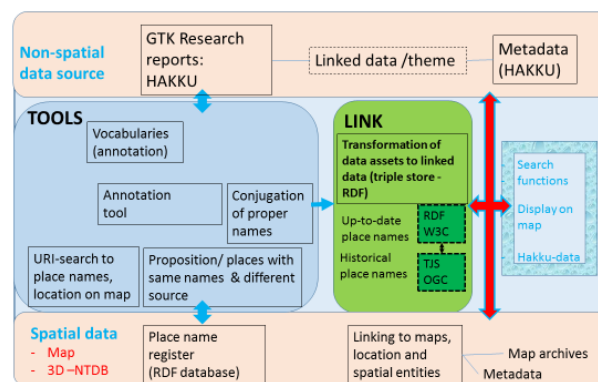


Figure 4: Georef architecture © Tiainen

2.3 Assessing the importance of named places: benefits and difficulties, Dominique Laurent, IGN France

Dominique Laurent from IGN France and EuroGeographics KEN INSPIRE presented current reflections and results of the United Nation Global Geospatial Information Management (UN GGIM) working group on Geographical Names. Named places refer either to features of well identified nature and geometry that have a name (administrative units, rivers, lakes, roads, streets, junctions) or to features that have a name but with a fuzzy geometry, possibly a fuzzy nature: populated places, land forms, forests, sea related feature. Named places are used for two main purposes: as search criteria (e.g. in gazetteer, GeoPortals) and for mapping. The first use case requires data completeness (users willing to find the named place associated with any geographical name) whereas the second use case require selection criteria (as it is frequently impossible to display all names and named places in the limited extent of a paper sheet or of a map screen).

In a first step, mapping agencies have selected relevant named places for maps at some given scale(s), following a cartographic viewpoint. However, this selection is very specific, both to a territory (and so difficult to harmonize across Europe) and to a scale or limited set of scales.



Figure 5: Designing a map requires selecting place names and deciding how to portray them depending on their importance but also on the cartographic context which includes the density of information and the length of the name.
(sources: IGN Top100 and Top25 maps)

In a second step, one of the objectives of the INSPIRE Directive is to make existing data interoperable. However, regarding theme Geographical Names, the data specifications have just included attributes about the least and more detailed viewing resolution, without any guidelines about how to interpret these subjective cartographic notions.

«featureType» NamedPlace	«codeList» NamedPlaceTypeValue
+ geometry :GM_Object	+ administrativeUnit
+ inspireId :Identifier	+ building
+ name :GeographicalName [1..*]	+ hydrography
«voidable, lifeCycleInfo»	+ landcover
+ beginLifespanVersion :DateTime	+ landform
+ endLifespanVersion :DateTime [0..1]	+ populatedPlace
«voidable»	+ protectedSite
+ leastDetailedViewingResolution :MD_Resolution [0..1]	+ transportNetwork
+ localType :LocalisedCharacterString [1..*]	+ other
+ mostDetailedViewingResolution :MD_Resolution [0..1]	
+ relatedSpatialObject :Identifier [0..*]	
+ type :NamedPlaceTypeValue [1..*]	

Figure 6: INSPIRE conceptual model for NamedPlace. A named place has several attributes relating to its importance. However, these are hardly reliable because subject to interpretation and in practice often void.

This is why, in a third step, the UN-GGIM: Europe Working Group on core data is proposing a more objective approach, by encouraging the estimation of the importance of the named place in the real-world, following a topographic, database viewpoint. The “Recommendation for content – Spatial Core data theme GeographicalNames” document is promoting the capture of quantifiable criteria measuring the importance of the named place in real world, such as its area (by capturing “true” geometry) or its population (for populated places).

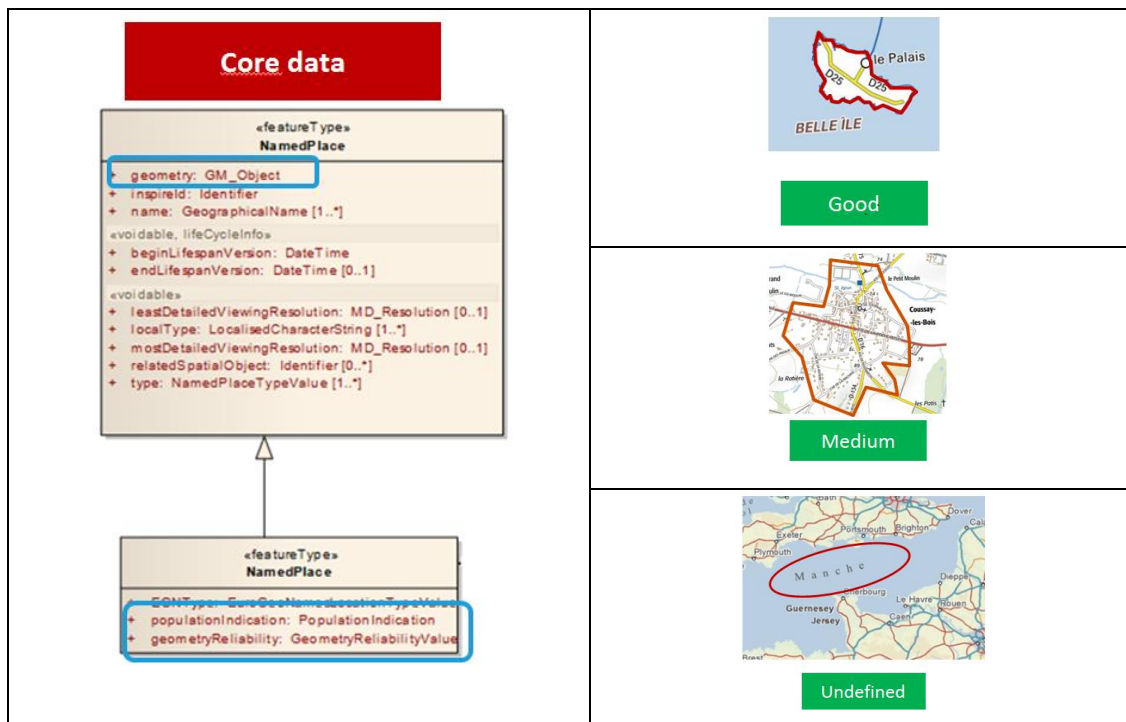


Figure 7: UN GGIM core data model for named place. A named place should be captured with its true geometry and its population (for populated places); in addition, the geometry reliability should be documented.

© Laurent. Sources of the maps: IGN Top100, Top 25, web.

However, there are some remaining issues and questions to be addressed, possibly by the research community. Capturing the “true” geometry of named places is both a challenge (how to do it in a reliable way whereas many named places, such as mountain chains or seas, have a fuzzy geometry?) and an opportunity (how much it could improve the linking by indirect spatial referencing?). The other potential research topic is related to the objective selection criteria: in addition to area and population, other criteria (e.g. touristic interest) have to be identified and methods of assessment have to be found.

2.4 Designing Data projects, how to value geographical heritage data with state of the art solutions?, Julien Homo, Kévin Darty, Foxcub

Searching digital assets more efficiently is a recurring need for many organisations even among their own assets. Julien Homo from Foxcub presented the Navigae project aiming at facilitating the community of geographers searching data produced by different research works, aerial pictures and maps, based on the spatial dimension of data, and explore the

context of these data thanks to interconnections with Wikipedia as well as rendering the geographical context (satellite views, old maps). The main criteria of search in Navigae are: space, time, scale. Navigae is built with state of the art geotagging Tools and web content. Once a resource is selected, Navigae display the links with Wikipedia as well as information extracted from Wikipedia, as shown on figure 8.

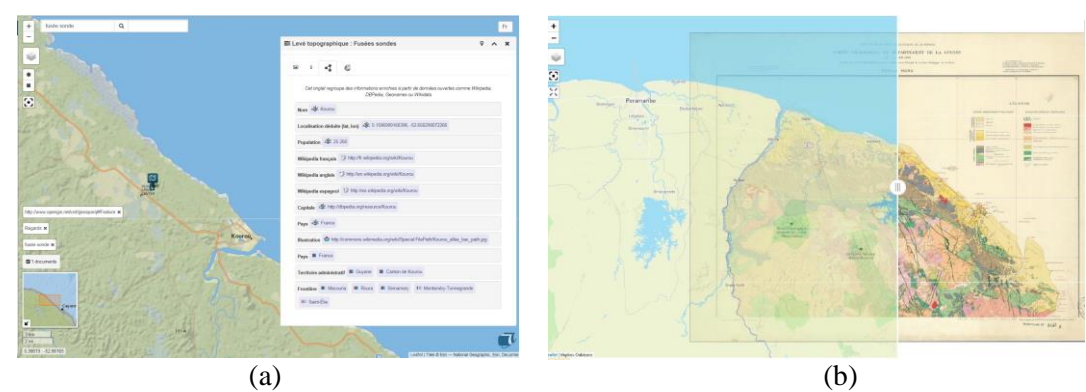


Figure 8: Snapshots from Navigae platform developed by Foxcub to locate geographical documents, to interconnect them with Web content like Wikipedia (a) and to support visual inter-comparison (b).

In such innovation project, Foxcub advocates the need for iterative and reactive approaches, to keep the Big Picture in mind and to be open to new technologies. Organising work to combine people skills rather than in silos is another learning from their projects. To achieve this, they developed a specific method grounded on a matrix usages x tasks (figure 9) which helps to make the right decisions at the right time in a project development.










Usages	Costs	Technical needs	Linking Data	Innovation	Tasks
					
Easily add new data sources	No licence		-	-	Done
Geolocate map with metadata	No licence	Need Machine Learning tool	Geonames Wikidata	Use Named Entity Recognition	Compare state of the art NER tools
Expose geographical data to the LinkedOpenData (LOD)	No licence		Geonames Wikidata DBPedia	Use GeoSPARQL	Done
Expose geographical data to geoserver	No licence		-	-	Done

Figure 9: Foxcub matrix to select the most relevant technologies in the course of the project
© Foxcub

2.5 Finnish Linked Data pilots, Kai Koistinen, National Land Survey of Finland

Kai Koistinen from National Land Survey, Finland, gave an overview and some live demos on Linked geospatial data pilots implemented in Finland.

A first proof of concept was implemented on spatial data identifier and URI service implementation, using the pattern [http://paikkatiedot.fi/so/\[datasetID\]/\[localID\]](http://paikkatiedot.fi/so/[datasetID]/[localID]). URI services are developed on top of Web Feature Services distribute: http data cards for human viewers (see figure 9) and machine readable format for applications. Both categories of http cards were indexed by Google.

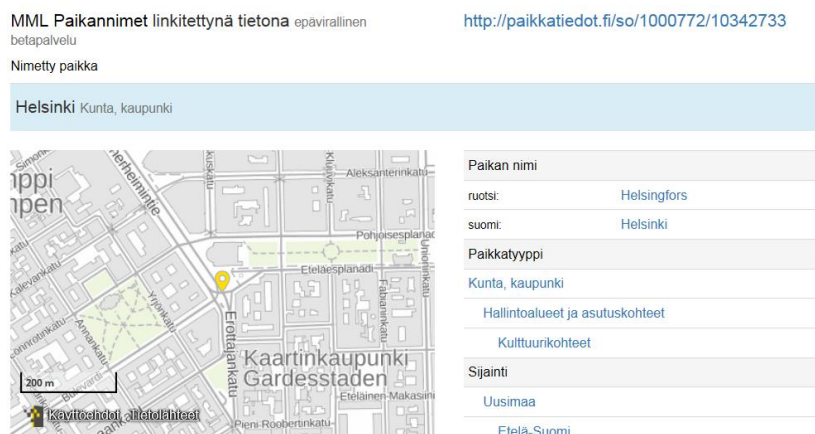


Figure 10: A http data card from the Finnish pilot © Kostinen

Links are also available in the description that connect named place with geographical names as well as named place with a sub-datasets related to this place, e.g. named buildings of Helsinki.

Another pilot addressed the need to build a national platform collecting data related to buildings from various public administrations. In the renewed topographic database, every building has a persistent http URI. Automatic scripts have been designed to generate RDF data from the relation spatial database. HTML cards were built on top of the Linked Data service – and not on top of WFS like in the geographical names pilot. The buildings data were linked to Wikipedia and Wikidata.

A third pilot was the integration of geographical data and areal classifications as Linked Open Data project (IGALOD). The aim was to connect Statistics Finland's areal classifications with NLS geometry data using Linked Data techniques. Both organizations provide SPARQL APIs for their data. The Municipality IDs, which appear in both datasets, are utilized to create URI links between the datasets.

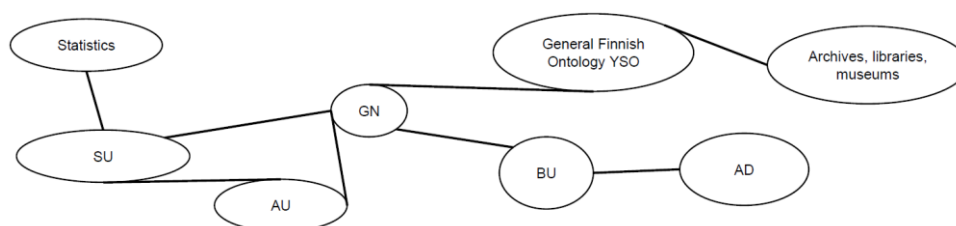


Figure 11: Geographical names are a glue between different datasets in Finland.
 SU: statistical unit, AU: administrative unit, GN: geographical name, BU: building,
 AD: addresses © Kostinen

As a conclusion from these pilots, GN theme is the glue between many themes, as shown on figure 11. Persistent http URIs and common ontologies are vital for interconnecting datasets.

2.6 The challenge of linking or integrating data on Buildings, Dominique Laurent, IGN France

In many countries, data describing buildings are scattered between different data producers and different products, to name but a few cadastral, mapping or statistical agencies products, as well different information systems like Housing Ministry or local governments. Besides, there are lots of documents associated to buildings, such as building permits, energy performance assessment reports, evacuation plans. In this context, providing an unifying framework to use building as indirect spatial referencing systems meets important needs. Most users would like to get access to the available information in an easy way, either by information of interest being integrated in a single data set or by information of interest being linked to reference geometric representation(s).

	Harmonisation relevant at European level			Harmonisation relevant at national/local level	
	INSPIRE Directive/GCM	International use cases	European Directives/initiatives	European Directives/initiatives	Local use cases
Widely available		Building/BuildingPart		OtherConstruction	Installation
	inspireId	heightAboveGround	numberOfFloors	constructionNature	installationNature
			currentUse		BuildingUnit
	beginLifespanVersion	elevation	dateOfConstruction	Association to CP	officialArea
	endLifespanVersion		dateOfRenovation		officialValue
	externalReference		conditionOfConstruction		address
			dateOfDemolition		
		buildingNature	numberOfDwellings		connectionToGas
	name		numberOfBuildingUnits		connectionToWater
					connectionToSewage
Rarely available			heightBelowGround		connectionToElectricity
			numberOfFloorsBelowGround		
			materialOfStructure		document
			materialOfRoof	Wall – Roof - Ground	floorDescription
			materialOfFacade		
			roofType	Opening	Room
			heatingSource/System		Texture
			energyPerformance		

Figure 12: Categories of information related to Buildings in the INSPIRE data specifications that might be considered in a unifying framework. © Laurent

However, this may be quite difficult to achieve due to the fact that there is no a single view on buildings: the same real-world entity may be considered as various features according various stakeholders, i.e. data producers will likely use different geometric representations and even different segmentations of buildings. For instance, the CityGML standard doesn't provide any clear guidelines about use of the Building and Building-Part concepts.

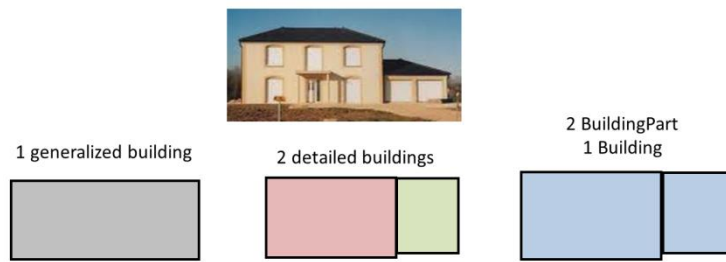


Figure 13: different views (segmentations) on buildings, even in a simple case © Laurent

The benefits and difficulties of integrating or linking data from various products on Building theme have been identified by several initiatives, such as the UN-GGIM: Europe Working Group on core data or the French Working Group on unique identification of Buildings.

Interconnecting pieces of different models across different perspectives requires:

- aligning the models (cf figure 12).
- handling inconsistencies in the way data producers implement models (segmentation of buildings).
- defining which data (attribute usually) should be documented in priority to achieve products with good semantic coverages.
- achieving coordination between data providers to offer a single easy access to INSPIRE core data related to buildings – instead of having to gather them

Dominique Laurent from IGN France and EuroGeographics KEN INSPIRE identified potential challenges for researchers from the current reflections of these groups. Research may be required to investigate both organizational issues (how to ensure efficient cooperation between various data producers?) and technical issues (which are the most frequent segmentation practices? which linkage mechanisms, e.g. address or unique identification of buildings, are the most efficient?).

2.7 Administrative Units as Linked Open Data – A case study from the Norwegian Mapping Authority, Thomas Ellett, Kartverket

In 2017, the Norwegian Mapping Authority, Kartverket, started work on a Linked Open Data project to distribute Administrative Units data through the RDF framework. The specific use case was to store administrative unit values in DCAT metadata as URI's, thus enabling better consistency of data, better handling of versioning and additional information made available to the end user through http URI's. The whole project has been completed using open source software and libraries, from Protégé with an Ontop plugin for ontology development and data transformation, to Virtuoso for RDF data storage and OpenApi and the Linked Data Theatre for data access endpoints.

The project planned to tackle four key areas: (1) development of a URI pattern, (2) creation of an Ontology, (3) transforming data to RDF and (4) delivery of that RDF data through multiple endpoints. Kartverket had some crucial elements already in place at the start; persistent local ID's at the object level, a stable UML model and data stored in a PostgreSQL database.

Thomas Ellett's presentation covered some basic theory on Linked Open Data and RDF, before delivering information about the technical elements of the project, both successes and challenges. He then presented information on how the general infrastructure has been setup and gave a live demonstration of the different endpoints available.

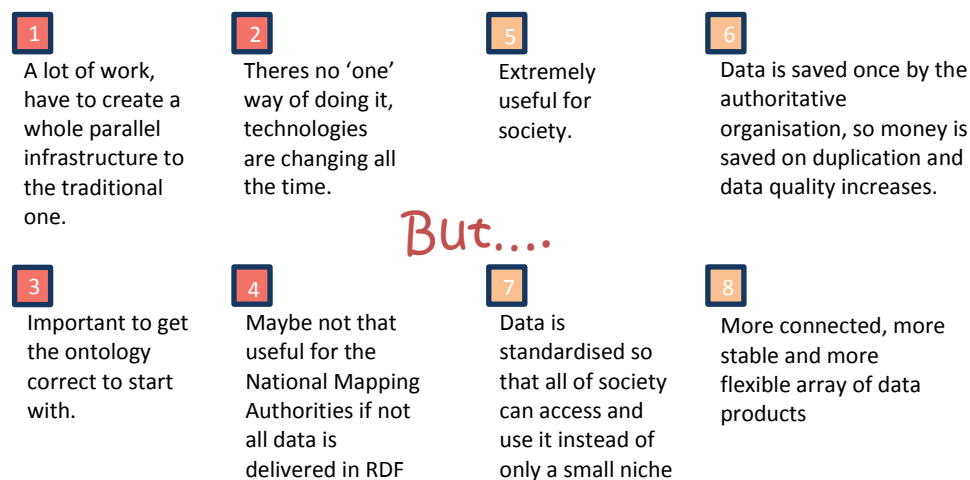


Figure 14: Pros (5 to 8) and cons (1 to 4) of Linked Data, lessons learned after Kartverkets pilot. © Ellett

2.8 Wikidata, a short introduction, Julien Boissel, Wikimedia foundation

Wikidata is a free collaborative multilingual secondary database dedicated to provide support for the wikis of the Wikimedia movement and to anyone in the world. Wikidata content can be queried from a SPARQL node, query.wikidata.org.

Wikimedia movement refers to the community of contributors to collaborative content projects supported by the Wikimedia Foundation (see figure 15) and contributors to a software that powers these collaborative content projects, called mediawiki. The Wikimedia Foundation is an American not for profit organization created in 2003 as a way to fund Wikipedia. There exist national chapters, not-for-profit organizations too, like Wikimédia France for instance.



Figure 15: Projects belonging to the Wikimedia movement comprise content project (Wikipedia, Wiktionary, Wikibooks, Wikiquote, Wikivoyage, Wikisource, Wikimedia Commons, Wikispecies, Wikinews, Wikiversity, Wikidata) and infrastructure and coordination projects (Meta-Wiki, Wikimedia Incubator). © Wikimedia

Wikidata structure is organized around items. An item represents roughly a page in Wikipedia (see figure 16). It has a unique identifier prefixed with Q. An item is described through statements composed of properties and values. Qualifiers are added on top of a statement to refine the statement scope if required – it can be seen as a metadata of the statement.

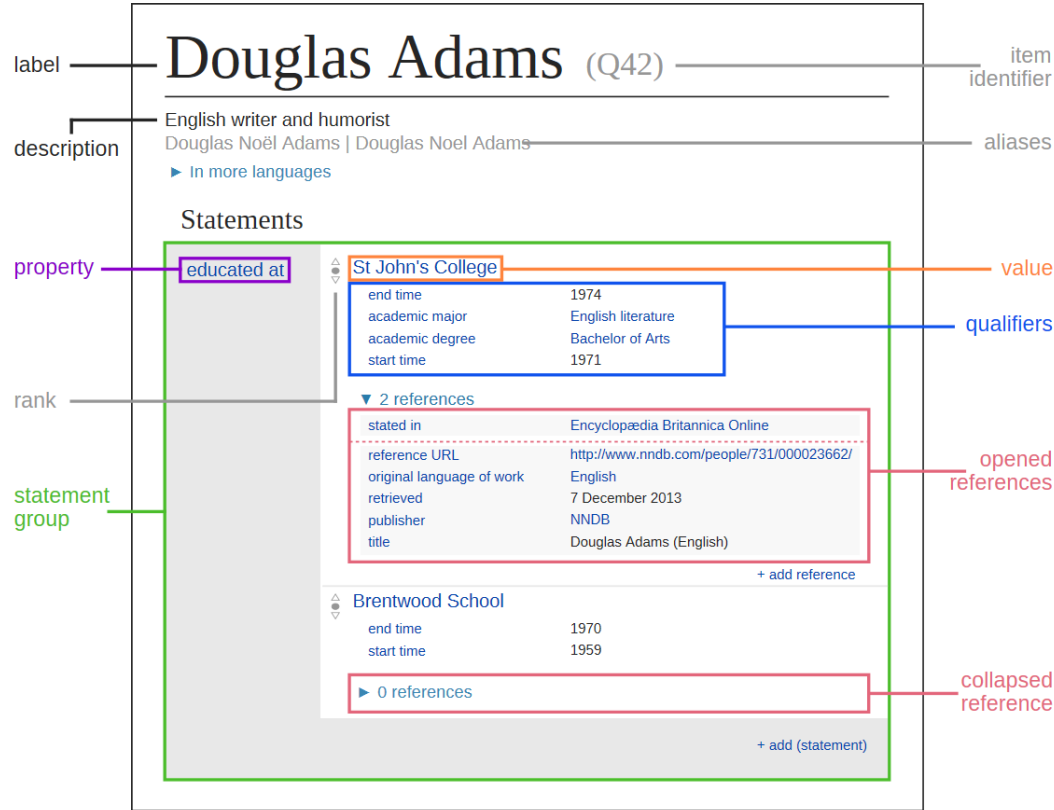


Figure 16: The item Douglas Adams in Wikidata. Source: Wikimedia, © CCO

In Wikidata, location may be provided either through the 'location' property with the type geo-location that is a Longitude and Latitude, or through indirect referencing systems, like the 'in' property that can associate a resource with a named place.

As explained by Sylvain Boissel, from Wikimedia France, specific curating methods are needed to achieve a trustable product. The curator watches for instance if features that are of similar importance in reality have similar levels of descriptions in Wikidata.

2.9 Linear indirect reference systems to interconnect data in transportation applications, Alain Chaumet, ENSG-Valilab

Location referencing practices are issued from two main domains.

The first domain roughly is maintaining order in a given area. It has led to the development of common techniques for military mapping and cadastre. First examples are the

Mediterranean maps, world maps and local descriptions, based on text or on drawing, of the agricultural land. Location practices developed from this perspective are geodetic networks and maps with national coverage.

The second domain is transportation and routing. This domain requires information about points of interest, towns with their specific designations and the distances between these locations. This graph of measured distances between known locations constitutes an indirect location system called linear referencing. The first known example of this kind of location systems is the Peutinger table which must have been very useful for travellers of the antique world. The standard for the interchange of geographic information related to roads is the GDF5.0 (ISO 14825:2011). And current CEN TC 287 focuses on Intelligent Transport Systems standard (2017).

In the domain of linear referencing, advantages are: the low segmentation of linear elements, the stable segmentation and real time information transfer. Disadvantages are: few easy to use applications, the computing of linear references depends on the used databases hence transfer of information from one database to another is not easy.

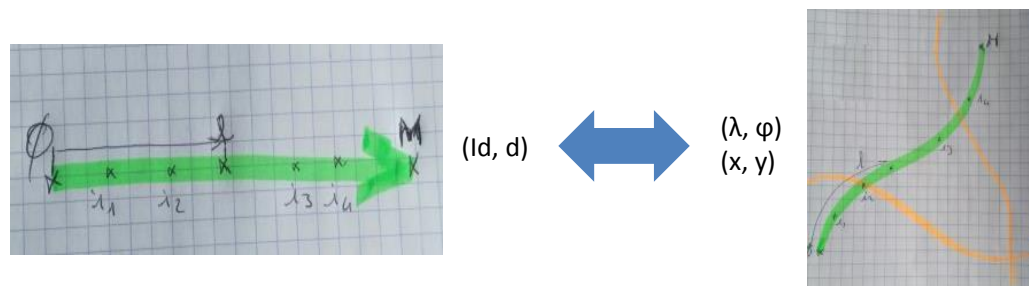


Figure 17: Reversible transformation from linear location to geographic coordinates spatial location (direct location) © Chaumet

Recent technologies should allow to merge both domains and develop consistent and easy to use translation application between direct reference systems (λ, ϕ or X, Y) and indirect systems like linear referencing commonly used.

Alain Chaumet then listed the expected benefits of merging direct and indirect location systems in the context of two running projects, the European EU-EIP (European ITS Platform) and the French LaSDIM (Large Scale Data Information for Mobility). He also pointed out the related issues to be solved:

- design easy to use tools for reversible transformation from linear location to (x, y) or (λ, ϕ)
- management of levels of details in road databases
- homogeneous linear referencing at various LoD (Levels of Detail)
- link between map makers and other road infrastructure stakeholders, use of LRS by NMCA
- daily or real time updating ; management of updating process from on ground events to data consumer.

3 COMMON FINDINGS ISSUED FROM WRAP UP SESSIONS

This section summarizes discussions and common findings: the need for specific ontologies, the need to develop new approaches to handle consistency in an open world, links management, communities management.

3.1 *Need for ontologies of places and of digital assets*

To make a better use of indirect spatial referencing for the management of growing digital assets, two categories of knowledge should be better represented in digital information infrastructures: shared knowledge related to places and knowledge related to digital assets themselves. This knowledge is referred to as ontologies in the following.

3.1.1 Need for ontologies of places

Ontologies of places could improve the detection and correct interpretation of an indirect spatial reference by a machine. This process is one of human reasoning currently difficult to reproduce by a machine. These ontologies should describe what is a place, how to model its geometry and its importance in a given context. In the domain of named places, they should also acknowledge places as social constructions. Time and scale should also be addressed because space is useful to correlate information across time or to change scales. Last, that one unique ontology can hardly be shared by all relevant actors and a more pragmatic objective, yet ambitious, aim at aligning ontologies and comparing how they describe places. Top level ontologies can be used to federate existing ontologies. These top level ontologies can concentrate on high level categories or on very generic descriptors. Another possibility is to design alignments between concepts of existing ontologies. Content specifications of national topographic maps, when written in some standardized language, can be seen for instance as such local ontologies about physical characteristics of named places. The structures of Wikipedia articles describing places also can be considered as candidate ontologies. The corresponding data has to be as complete as possible (to enable as many links as possible), reliable (e.g. correct spelling of names to ensure correct links), rich or connected (with enough information to enable disambiguation of geographic identifiers, if several ones are similar).

3.1.2 Need for ontologies of digital assets

Ontologies of digital assets also are required to facilitate our study of digital assets –e.g. the content produced on social networks, or the digitized archives of a public administration, etc. They are important to help developers grab the complexity of current digital products that are very often derived and assembled from different sources, and different technologies, which make the assessment of their accuracy difficult. They are also needed to facilitate the way machines or developers assess the scope of an information asset, be it an atomic RDF statement or a data set or a data service. Scope is often explicated through the following metadata: spatial and temporal coverage, content, provider, lineage information, quality criteria and possible usages. In indirect referencing, the notion of scope can firstly refer to the scope of a corpus: what kind of indirect references are expected to be found in it. It can also depend on the form of every item, like for instance administrative documents. Knowing this scope is important to be able to train tools that are used to reference the corpus. It can also refer to the scope of gazetteers: what kind of place names and disambiguation information can be found in the gazetteer. Knowing this scope is important to select the right gazetteers during a “text to space” process.

3.2 *Consistency in an open world*

3.2.1 *Need for similarity and relevance measurements*

In today information infrastructure the open world assumption has replaced the traditional close world assumption of legal systems. Information is queried on the basis of its relevance to a given request as opposed to query system with exact criteria where searched information asset should satisfy exactly query criteria. Relevance measurements are becoming important as well as similarities between query contexts or similarities between queried assets because they support query extension, recommendation and assets browsing.

3.2.2 *Reconstructing minimal subsets with consistent scopes*

A specific challenge is to extract from a set of facts/statements/data the subset that fits a given scope of interest – in terms of space, time and theme.

3.2.3 *Adopting description logics, adopting LOD technologies*

Relevant technologies for handling an Open world information infrastructure are Description Logics and Linked Open Data (LOD). Yet, even if theories are mature like Description logics (OWL, ..), the software still is not mature enough and implementing a production system with large amount of data and full SPARQL support might be impossible with available solutions. LOD is a method of publishing structured data so that it can be interlinked and reused from different sources. LOD builds upon standard Web technologies - HTTP, RDF and URIs-, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. (https://en.wikipedia.org/wiki/Linked_data). We need consistent implementation methodologies for LOD delivery to ensure enough interoperability. When it comes to designing and distributing indirect referencing systems, workable off-the-shelf software, open source or commercial, are not available; co-development, even code development in some stages of data linking process as well as knowledge exchange is most necessary. The Joint Research Center is targeting the sharing of good practices with the Spatial data on the Web report.

3.3 *Computing, maintaining and sharing links*

Data linking can refer to pre-computing reusable links and publishing them, or to computing on the fly new links between contents. In our opinion, in both categories of processes, the definition of ontologies and clear commitments for curating is needed.

In traditional SDIs, the general hypothesis is that there are too many relationships to store them all, but for some topological relationships. In general, one can interconnect content based on co-visualisation or based on spatial data matching tools that process coordinates and other attributes. Co-visualisation is not always straightforward and pre-computing may be required to adapt the data and generate a consistent visualization layer. Data matching requires in theory chaining different Web Processing Services or in practice downloading data on one's computer, transforming them to be in the same data format and then applying spatial analysis tools.

Linked data proposes a way to explicitly represent relationships – as opposed to coordinates. This was also the motivation of the Table Joining Service (TJS) working group of OGC. In this respect, linked data are promising because the community agrees on what the relevant relationships are that need to be ‘shared’, rather than computed on the fly. With respect to

traditional SDIs, publishing relationships is especially relevant when they involve semantics which coordinates cannot represent without complex spatial analysis methods. Besides, when it comes to computing interconnection on the fly, the SPARQL query of LOD technologies can express queries that search data from different servers (SPARQL nodes) which is especially relevant to interconnect content distributed from different authorities, hence different servers.

Commitments for curating are needed to maintain the links when interconnected content evolve.

A specific category of interconnection procedure is to transform a reference in one system into a reference in another system and not lose any precision. Here also, algorithms need a fine description of datasets, especially geometries, and how to interpret them.

3.4 *Communities, commitments, authorities*

Even though spatial capacities are ever more requested today, few data providers are willing to cope with these specific data and the corresponding quality management issues. This is why using references to authoritative data source, maintained by some legally mandated organization or trustable entity is targeted by more and more organizations.

The notion of authority is related to law and to usage. The law is not always very specific about the indirect spatial referencing framework that should be used. With respect to usage, it is interesting to look at the cloud of linked data where Wikipedia is at the center, meaning that it is a content mostly referenced to. Wikipedia is not specifically spatial, yet the Wikidata knowledge base is a valuable resource because of these unique characteristics:

- the very strong Wikimedia community and expertise in community management Wikipedia,
- the character of de facto standard of Wikipedia on the Web when it comes to making a reference to a description, to an explanation,
- the multilingual aspect of both Wikipedia and Wikidata is a native feature especially interesting to connect to people, including people who can contribute to ontologies and to data.

Even if national mapping agencies and other content providers have begun working towards maintaining reference databases, they have not yet committed to implementing http URIs which would be the access point to this information. This lack of commitment can be explained by uncertainties related to the benefits and costs of this technology.

Benefits from adopting linked data listed by participants are:

- Financial benefits through reduction in data duplication.
- Increased efficiency in research extension and report generation.
- Increase in quality of reference data as it won't be downgraded over time through copying and manual editions.
- Increased quality of thematic data through better common semantic understanding and better semantic consistency across domains.
- Possibility of data driven applications, utilizing the RDF framework and its expressiveness to exchange information update between applications (as opposed as delivering the whole geo-dataset to the application).
- Adoption of our data services by new users who are developers of application consuming Web content.

Costs related to adopting linked data listed by participants are:

- Financial costs arising from development of separate LOD infrastructures.
- Financial costs related to R&D, since the methodologies currently available for the production and consumption of LOD, and application of ISR are not mature. However Finnish pilots show that transforming the data to RDF and generating individual HTML-pages of data objects is cost-effective with self-made scripts where a common practice and guidelines could be developed.
- Increased vulnerability of reference data that is stored just once, which will probably lead to substantial financial and resource costs.

4 CONCLUSION AND FURTHER AREAS OF RESEARCH AND DEVELOPMENTS

A follow up of the seminar is the identification of challenges that can be shared with scientists and developers in the domain of Indirect Spatial Referencing, mainly related to two communities: text to space and linked data. Several categories of challenges are identified and work will go on to implement them within EuroSDR.

1) Ontology design, knowledge representation

- Place ontology design: cataloguing, exploring common ontologies (incl. Wikidata). Modelling named place types to address heterogeneities in national gazetteers.
- Ontologies for digital assets:
 - o New concepts to replace the notion of data product, data series, data set.
 - o Propose more semantics to links, that can also be useful for deciding if the links must be pre-computed or computed on the fly, and for the maintenance of the links.

2) Alignment, interconnection:

- Alignments of place ontologies
- Automatic interconnection of schemas – incl. INSPIRE specifications, incl. OSM tags
- Automatic interconnection of data related to buildings
- Connecting INSPIRE data with important web content like Wikidata, Wikipedia, OSM
- Updating links
- Migration methods between linear and absolute referencing

3) Relevance, similarity, ranking

- Assessing relative importance to named places: the challenge consists in documenting a specific property for a place that: its importance in a given context. This property will be used to choose how to portray it in a map when zooming in and out.
- Adapt a datacard, i.e. the set of information displayed when someone queries a given URI, to a context: portray only relevant metadata, recommend relevant additional resources
- Proposing similarity measures between places

4) Technology readiness assessment in NMCA context:

- Benchmarking Linked Data software with large amount of data and full SPARQL support
- Benchmark various practices of linear referencing, developing more or better methodologies and tools to migrate data from one system to another one)

This organization of challenges should be consolidated with the Joint research centre which hosts reports and registries on best practices.

Another conclusion was the importance of connecting people within NMCA environment who are working with new technologies – text to space, linked data – and can be isolated. A solution could be to create technical groups that can act as focus forums with a light investment of participants who often are developers with little time to devote to networking.

5 ACKNOWLEDGEMENTS

The workshop was supported by EuroSDR. The authors wish to thank Joep Crompvoet from KU Leuven and François Golay from EPFL for their careful and constructive review of a first version of this report.

ANNEX 1 – PROGRAM

Day 1: September 4th – 13:30-18:00

13:15-13:30: Registration, welcome coffee

General introduction,
Bénédicte Bucher, IGN-France

Georef, Service and Development Platform,
Esa Tiainen, National Land Survey Finland

Gazetteers for linking text to space: experiences with contrasting corpora,
Elise Acheson, University of Zurich, Switzerland

Designing data projects, how to value geographical heritage data with state of the art solutions? Julien Homo, Kévin Darty, Foxcub, France

Assessing the importance of named places: benefits and difficulties,
Dominique Laurent, IGN France

Discussion on data linking by place names

Day 2: September 5th – 9:00-15:00

Finnish Linked Data pilots,
Kai Koistinen, National Land Survey, Finland

The challenge of linking or integrating data on Buildings,
Dominique Laurent, IGN France

Administrative Units as Linked Open Data – A casestudy from the Norwegian Mapping Authority, Thomas Ellett, Kartverket, Norway

Wikidata, a short introduction,
Julien Boissel, Wikimedia foundation, France

Linear indirect reference systems to interconnect data in transportation applications,
Alain Chaumet, ENSG-Valilab

Lunch

Wrap up, drafting position papers and challenges