Land Governance Lost in Translation - Exploring Semantic Technologies to Increase Discoverability of New Technologies & Data

Lisette MEY, Netherlands and Laura MEGGIOLARO, Italy

Key words: controlled vocabulary, semantic web, semantics, discoverability, knowledge sharing, data exchange

SUMMARY

Language and technology barriers are a very serious constraint to effectively exchange and learn from land data, information and technologies across the globe. We would like to explore whether we can gain inspiration from how semantic web technologies have overcome knowledge-sharing challenges in other sectors, such as the agriculture sector. With emerging technologies, new tools and ever-growing amounts of land data, we face a very real risk of losing the overview. Without this overview, data is much less likely to be used and thus be useful. We will particularly look at the use and value of controlled vocabularies for the land sector.

Land is a topic that is debated in many languages, across different (academic) disciplines and in all parts of the world. Furthering our collective agenda, sharing and learning from knowledge and perspectives from other contexts, or transitioning technological innovations from one country to the other is complicated by - among many other aspects - language and terminology barriers. Many attempts have been made in the past to find common definitions and terminologies for issues related to land, but a wide consensus or adoption has never been reached. Understandably so: one can only imagine the heated and controversial discussion to reach agreement on what we mean exactly when we use the word ‘property’. It simply does not have the same meaning in each country or context. It is a daunting and arguably impossible task to reach this global consensus. In this paper, we will present our experience with controlled vocabularies and the opportunities and challenges it can bring.

Land Governance Lost in Translation - Exploring Semantic Technologies to Increase Discoverability of New Technologies & Data

Lisette MEY, Netherlands and Laura MEGGIOLARO, Italy

1. INTRODUCTION

Language and technology barriers are a very serious constraint to effectively exchange and learn from land data, information and technologies across the globe. We would like to explore whether we can gain inspiration from how semantic web technologies have overcome knowledge-sharing challenges in other sectors, such as the agriculture sector. With emerging technologies, new tools and ever-growing amounts of land data, we face a very real risk of losing the overview. Without this overview, data is much less likely to be used and thus be useful. We will particularly look at the use and value of controlled vocabularies for the land sector.

Land is a topic that is debated in many languages, across different (academic) disciplines and in all parts of the world. Furthering our collective agenda, sharing and learning from knowledge and perspectives from other contexts, or transitioning technological innovations from one country to the other is complicated by - among many other aspects - language and terminology barriers. Many attempts have been made in the past to find common definitions and terminologies for issues related to land, but a wide consensus or adoption has never been reached. Understandably so: one can only imagine the heated and controversial discussion to reach agreement on what we mean exactly when we use the word 'property'. It simply does not have the same meaning in each country or context. It is a daunting and arguably impossible task to reach this global consensus. In this paper, we will present our experience with controlled vocabularies and the opportunities and challenges it can bring.

2. THE POTENTIAL OF THE SEMANTIC WEB

2.1 What is the semantic web?

Tim Berners-Lee, the inventor of the world wide web, once described the semantic web as follows:

> "*I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A "Semantic Web", which makes this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be*

handled by machines talking to machines. The "intelligent agents" people have touted for ages will finally materialize."[1]

There is a wealth of data and information available on the web, more being added every day from every part of the world. It has become impossible for humans to digest this all and be aware of all elements online. It is sometimes said ironically, that the answer to the world's problems lie in a PDF somewhere online. But someone needs to find, access and digest this information before being able to actually solve the world's problems. We would not want to go that far as saying "all the world's problems" can be solved with already existing information, but there is definitely truth in the fact that we can benefit more from existing knowledge and tools to address issues that happen globally.

Generally, new technologies (for example, on data capture or innovative surveying methods) or newly generated knowledge are shared among personal networks, such as the FIG network. But what about people that do not have access to such networks? Knowledge remains confined within certain siloes, whether they are thematic (land administration vs. gender experts, for example), sectorial (surveyors vs. grassroots activists, for example) or geographical. Not accessing all potential beneficial knowledge and tools is therefore partially an issue of breaking out of old habits, but even if the will was there - where do you possibly begin? If a simple Google search for 'surveying techniques' returns over 34 million records, even the best intentions are not going to be enough. It is simply too much for a human to digest this wealth of information.

The semantic web aims to address just this. The goal of the 'semantic web' is to make information available online machine-readable. Humans cannot digest all this data and information, meaning that important knowledge will never reach its full potential or even, in the worst case scenario, remain unused. Machines can help us read and digest this information at an unprecedented speed or scale. In order to effectively share knowledge and technologies across the globe and increase our collective efficiency - we need to embrace a tool like the semantic web.

2.2. **What is machine readability?**

To understand how we can embrace the semantic web as a tool for effective knowledge sharing globally, we need to understand what machine readability is. The common perception that anything put on the web can be read by machines, is woefully incorrect. It is true that many applications or software instances have been developed to digest more and diverse types of information, such as pictures, PDFs or even satellite images. But such applications are often very expensive to develop and perfect, and as such, as hardly ever affordable for non-commercial organizations to use. Particularly when we consider people and organizations working in less developed countries. The idea of the semantic web does not envision 'machine readability' through applications or software, but rather non-proprietary machine readability.

[1] Berners-Lee, Tim; Fischetti, Mark (1999). *Weaving the Web*. HarperSanFrancisco. chapter 12.

Important to remember is that machines read in 0s and 1s, and therefore structure, standards and formats are incredibly important for a machine to fully understand the meaning of data or information. The semantic web is based on 'Resource Description Framework' (RDF) which is a machine-readable technology based on triples: object, predicate and subject.[2] Structuring information, particularly metadata, in such a way allows machines to understand what it is about and help retrieve information to an end user. This may sound convoluted, but it is something anyone that has uploaded any information to a repository, has dealt with.

> *Think of a simple example of uploading a paper to an online library or journal. You will be required to fill in certain fields describing your paper. The 'object' (first of the triples) you are describing is: your paper. The 'predicates' (second of the triples) are the different fields that you are required to fill in. A title-field, for example, will have "hastitle" as predicate in the backend of the online library. The subject (third of the triples) is the actual title of your publication. A machine will read: "your paper" >> hastitle >> "title".*

Three elements are of crucial importance in the back end to make this information machine-readable: format, uniqueness and standards.

2.2.1. Format

Firstly, the format needs to be open. As mentioned before, for a machine to read PDF or an Excel file, it will need programs such as Adobe or Microsoft Excel. The principle of machine readability is that such proprietary software will *not* be needed. This RDF-based metadata therefore should be in a format such as CSV, JSON or other open-formats. We will not go into this topic of formats much deeper, because much has been written on the topic.

2.2.2 Uniqueness

Secondly, uniqueness is very important. Remember that machines read in 0s and 1s, therefore the title of a paper such as "New Surveying Methods" is read as a combination of certain 0s and 1s. Another paper with an exact same title, will have the same combination of 0s and 1s. Or if we are talking about the name of a tool for example, this may change over time. How will the machine be able to understand that papers with the same name, are in fact two different papers (and how will it attribute the right RDF information to the right paper)? Or how will a machine know that the two names the same tool has had over time, are in fact the same tool?

A machine will need to be able to differentiate. This is why in the semantic web, the use of unique IDs is of crucial importance. Think of how papers in journals often have a DOI-number or published books have an ISBN-number. The same should go for resources published on the (semantic) web: resources should have a unique ID to ensure that machines will always be able to attribute meta-information about this content to the correct and unique resource.

[2] World Wide Web Consortium (W3C), "*RDF/XML Syntax Specification (Revised)*", 10 Feb. 2004.

2.2.3. Standards

A third crucial element to machine readability is standards. Take the example we mentioned above: how does a machine know that the "hastitle"-predicate is actually a *title* of an object? Because the predicate is based on a standard. Standards have been developed for metadata, formats, data structures -- all in a way that machines are able to understand them. We can write hundreds of papers and probably several PhD-studies can be conducted digging into the different standards, how they work and how they were developed. In this paper we want to focus on one type of standard in particular: controlled vocabularies.

2.3 What are controlled vocabularies?

A controlled vocabulary, in short, provides a way to search and discover data and information. Controlled vocabularies are used in libraries, repositories and any other knowledge storage system for indexing information.[3] The concepts in such a controlled vocabulary are used to tag data and information. Using a controlled list of concepts, issues such as synonyms, homographs or translations are circumvented. It is, in other words, a standard for keywords.

This is another critical element for the effectiveness of the semantic web. If a user queries a database, for a machine to be able to retrieve relevant information, it is important that the computer also understands what the topic is. If anyone can fill in anything when they upload content to this database, the machine has no way of knowing relationships between terms of how a resource tagged with a synonym, might also be of interest to this user.

Controlled vocabularies work with unique IDs for each concept, with the possibility of adding several labels to that ID: the preferred term, translations in an endless number of languages, relationships between terms (A *is related to* B, or X *influences* Y, etc.). This way the machine can understand the languages and the nuances we use in languages, and help retrieve the most relevant and to-the-point information to a user's query. We will dive deeper into the potential of controlled vocabularies by highlighting the case of AGROVOC, the agriculture thesaurus.

3. THE CASE OF AGROVOC

AGROVOC is a controlled vocabulary established and facilitated by the Food and Agriculture Organization (FAO) of the United nations. It covers "*all areas of interest to the FAO, including food, nutrition, agriculture, fisheries, forestry, environment etc.*".[4] The AGROVOC thesaurus was first published (in English, Spanish and French) in the early 1980s. In 2000, AGROVOC went digital. It has evolved and grown over the years, with a vibrant and international community of editors behind it, contributing new concepts and new translations every month. Today, AGROVOC consists of over 36,000 concepts and over 750,000 terms (synonyms or translations to those concepts, etc.) related to agriculture and is translated to over 35 languages.

3

4 AIMS (2019), "*AGROVOC | Agricultural Information Management Standards*".

AGROVOC is widely used in specialized libraries as well as digital libraries and repositories to index content and for the purpose of text mining. It is also used as a specialized tagging resource for content organization by FAO and third-party stakeholders. FAO statistics show that the vocabulary is used by 1.8 million users every month to classify agriculture data and bibliographic resources. AGROVOC has thus increased the visibility and discoverability of agriculture data and information to an immeasurable scale.

A controlled vocabulary such as AGROVOC, has helped no less than 10 million users a year in overcoming the language barriers we just described. Through AGROVOC's technical infrastructure, computers can read concepts beyond 0s and 1s and understand how 'maize' as a concept is the same as 'Maïs' in French or 'ذرة صفراء' in Arabic. Translations, synonyms and relationships of this one concept are captured in one unique code, a 'Uniform Resource Identifier' (URI) , that computers, including search engines, can read and understand.

4. WHERE IS THE LAND SECTOR?

With such an incredible tool and even more incredible user base as AGROVOC, one quickly starts thinking: what about land? If the AGROVOC tool covers all areas of interest to the FAO, surely land governance must be one of the topics they cover. When the Land Portal Foundation first discovered AGROVOC and engaged with the team, only 20 concepts related to land governance were included in the AGROVOC vocabulary.

4.1 Gap exploration research in use of controlled vocabularies in land sector

As a part of the GODAN Action-consortium, in 2016 the Land Portal Foundation did a scoping study of land information providers online and the way they classified their information. Or in very simple words: what kind of tags do they use? The main conclusions about the use of standard vocabularies within the land governance community is that there is no structured or uniform approach to use them to publish information. We saw a range of sophistication in the way to classify the materials the organization publishes, starting from no classification at all, to a standard set of keywords that could be used.

Roughly, five types of classification were identified. The first being no classification at all for content or merely categorizing content by resource type (see for example the Asian Farmers Association's website). Secondly, many organizations use a 'free tagging'-system, allowing the users to create new tags as they add new resources (see for example the AgEcon website, maintained at the University of Minnesota by the Department of Applied Economics and University Libraries, and the Agricultural and Applied Economics Association), leading to an unstructured list of thousands of keywords that overlap. The third situation is where organizations have a standard set of keywords that can be used to classify content, but there is no real structure to these keyword lists. For example, organizations do not differentiate between resource type, geographical keywords or topical keywords within these lists (see for example the Asian NGO Coalition or the Focus on Land in Africa (FOLA)-website, a joint initiative of the World Resources Institute (WRI) and Landesa). Similarly, some organizations do have a standard set of keywords or topics, but that standard is only applicable to their own organizations and not meant to be re-used or accepted by other organizations. See for example

the International Land Coalition website, that has structured their publications under their own strategic commitments – that not even their partners, who as members of the Coalition have committed themselves to the same goals - have adopted on their own websites.

Finally, there have been attempts to standardize a set of topical keywords – a glossary - within the land sector and to gain general acceptance of the entire sector to these initiatives, such as Focus on Land in Africa (FOLA) and more recently, the Global Land Indicator Initiative (GLII). However, these glossaries are stand-alone lists in HTML or PDF format, but not used or applied in any way. Focus on Land in Africa (FOLA), as mentioned above, does not use their own glossary to classify their content – it is meant to merely guide users through the documents they can read on the website and to create an understanding behind the meaning of the different keywords. The Global Land Indicator Initiative has created a glossary with key land-related terms, which has been a collaborative process by several prominent organizations working on land. However, this list has not been published yet, nor are there any concrete plans to use this glossary other than as a reference for generally accepted and determined key concepts and definitions for land governance issues.

Conclusions from these different classifications within the land sector that were identified during the scoping research, is that there is a very limited awareness about standards to classify data within the land sector. Some organizations do not use topical keywords at all and those that do, have not designed these lists to be seen or used by other organizations at all. Therefore, there is a clear gap in the use of standards for the land sector and in the existence of standards for the land sector specifically.

5. INTRODUCING LANDVOC - THE LINKED LAND GOVERNANCE THESAURUS

The Land Portal Foundation has responded to this gap, not by creating yet another new standard, but by taking a widely accepted and used standard such as AGROVOC and enriching the concepts related to land within this vocabulary. By building on existing land glossaries, such as the FAO's Land Tenure Thesaurus (developed as a reference point for FAO staff), or the Land Administration Domain Model or the Global Land Indicators Initiative. New concepts were added and translated to several languages. This particular set of land-governance related concepts in AGROVOC is now called "LandVoc - the linked land governance thesaurus".

LandVoc can be an extremely powerful tool in making data and information more discoverable. It can connect knowledge and experiences from across the world, bridging both language and culture barriers. LandVoc is intended to be an unbranded linking tool between the different classification and tagging systems information providers in the land sector use.

5.1 Challenges

There is no doubt that the land community experiences the same struggles in language-differences as they do in agriculture -- however, arguably, these are much more nuanced and complex. With a topic such as land, classifications are controversial and immediately become

political. Furthermore, in a sector where multiple tenure systems coexist within one country (all with their own associated terminologies) and that harbors immense power imbalances between global and local, between government, private sector and local communities -- uttering the phrase ‘standardizing’ is often considered either naive or some sort of utopia we will never reach. In such discussions, we hear that land experts feel that acknowledging the differences in the way we choose to name or describe the issues we face, however evident or subtle these differences may be, has to be more important than increasing discoverability of information.

Enriching the land concepts in AGROVOC to try and capture the nuances of land governance in the LandVoc vocabulary goes beyond technical features, people tend to argue, but is something more fundamental: it is scientific, psychological and political in nature. We could not agree more. As a team whose everyday business involves managing an information technology platform, we cannot help but see the technological benefits of such a tool. But we also see that in global thesauri, English remains the dominant language and the starting point that other languages build on, rather than entering from their own perspective. We see that, when it comes to definitions or preferred terms to use, Western perspectives and interpretations of concepts are much more dominant than those of stakeholders in the global South.

In facilitating a standard vocabulary for land, our intention is not to counteract such differences or ‘impose’ a standard for a particular concept -- but rather, to build a tool that embraces and highlights our differences. Thus, providing a basis to gain a deeper understanding of the issues we deal with and how they vary from stakeholder to stakeholder and context to context. We are aware of the fact that we will never be able to capture all languages, nuances and differences, but, in our opinion, this isn’t a reason to not begin trying! We would argue it is actually quite important to realize and acknowledge that when a researcher that has a PhD with regards to a certain topic uses a certain term, it means something different than when a practitioner working at intergovernmental organization uses the same term. Currently, there is no way for a layman to realize this, other than by speaking to such stakeholders individually.

We have a choice: we can carry on conversations with those select few that understand and acknowledge our particular conceptualization of land governance and limit the outreach and impact of our work, or we can choose to be more inclusive and decide to embrace and convey these important differences to a wider public. If tools such as a Google search engine are used by millions of people already, LandVoc can help to ensure that others can also begin to gain an understanding of the rich complexity and controversy of a topic such a land governance.

5.2 Opportunities

Not only is the Land Portal Foundation active in the land sector to promote standardization and work constructively on making land data and information more discoverable - however daunting that task may be - the Land Portal is also a major advocate within the open data-community not to duplicate efforts or standards, but still make universal standards useful for smaller expert communities.

Of course AGROVOC largely overlaps with possible land concepts, but using solely the agriculture standard will not be relevant enough to meet the land sector's needs, because it also contains thousands of concepts that are not relevant to land. Recognizing that the overlap between the two standards would be significant and not wishing to duplicate efforts, the Land Portal and FAO explored options on how the AGROVOC thesaurus could be made useful to specific expert communities.

The solution brought forward and currently implemented, is that of the multi-hierarchy scheme. Land concepts will be in AGROVOC, within the AGROVOC hierarchy, but there will also be a separate scheme within AGROVOC, that only contains concepts related to land governance: "LandVoc". This LandVoc scheme can have its own independent hierarchy from AGROVOC. This solution allowed to avoid duplication of efforts, but still making the thesauri relevant for the specific expert communities. AGROVOC is now exploring these options for other expert communities as well, such as fisheries and soil.

With such a great infrastructure for a new tool as LandVoc, the Land Portal Foundation has performed a year-long consultation with experts building the independent hierarchy for LandVoc. This will make it an even more useful tool for the land sector to use.

6. CONCLUSION

We have seen how semantic technologies, and particularly the use of controlled vocabularies, can increase the discoverability of data and information considerably. AGROVOC, has increased the visibility of agriculture data and information and serves an audience of over 1.8 million users per month. Land Portal's research has shown that the land sector is far from reaching such a potential since no standards are being used to classify land data and information online.

The Land Portal saw this gap and worked with the AGROVOC team at FAO to increase the 20 land-related concepts in AGROVOC to 300 unique concepts, excluding the added translations and synonyms. This set of land-related concepts within AGROVOC is called "LandVoc". LandVoc could similarly increase the visibility of land data and information and help the way we exchange land data across the world. More than that, it can also serve as a reference document for translations and to capture and understand the richness and complexity of land governance terms.

REFERENCES

AIMS (2019), "*AGROVOC | Agricultural Information Management Standards*".

Berners-Lee, Tim; Fischetti, Mark (1999). *Weaving the Web*. HarperSanFrancisco. chapter 12.

World Wide Web Consortium (2004), "*RDF/XML Syntax Specification (Revised)*".

BIOGRAPHICAL NOTES

CONTACTS

Lisette Mey
Land Portal Foundation
Bakboord 35 3823TB
Amersfoort
THE NETHERLANDS
+31657710841
lisette.mey@landportal.org
www.landportal.org