Feasibility of Information Interoperability in the Humanitarian Domain

Tim Clark
Milcord
303 Wyman St, Ste 300
Waltham, MA - 02451
tclark@milcord.com

Carsten Keßler
Center for Advanced Research
of Spatial Information and
Department of Geography
Hunter College, CUNY
695 Park Avenue
New York, NY - 10065
carsten.kessler@hunter.cuny.edu

Hemant Purohit
Ohio Center of Excellence in
Knowledge-Enabled Computing (Kno.e.sis)
Dept. of Computer Science and Engineering
Wright State University
3640 Colonel Glenn Highway
Dayton, OH - 45435
hemant@knoesis.org

Abstract

Given the rise of humanitarian crises in the recent years, and adoption of multiple data sharing platforms in offline and online environments, it is increasingly challenging to collect, organize, clean, integrate, and analyze data in the humanitarian domain. On the other side, computer science has built efficient technologies to store, integrate and analyze structured data, however, their role in the humanitarian domain is yet to be shown. We present a case of how structured data technology, specifically Linked Open Data from the Semantic Web area, can be applied for information interoperability in the humanitarian domain. We present the domain-specific challenges, description of the technology adoption via an example of real world adoption of the Humanitarian Exchange Language (HXL) ontology, and describe the lessons from that to build the case of why, how and which components of technologies can be effective for information organization and interoperability in the humanitarian domain.

Introduction

The humanitarian domain has been evolving with an unprecedented level of heterogeneous data in the recent years. The amount of data to manage has exploded in face of new data sourcing and sharing mechanisms, and especially with the growing number and scale of crises all over the world. Over the past three decades, the frequency of natural disasters has increased multifold globally (EM-DAT 2014). Moreover, the economic losses of hundreds of billions are worrisome (ReliefWeb 2013; UNESCAP 2013), presenting a case for a framework of data-driven response planning. Such an ambitious goal of a data-driven approach to organize, aggregate and analyze the vast amount of heterogeneous data requires investigation into how existing and adaptable structured data technologies can be adopted in this domain.

Existing practices for data management in the humanitarian domain include a lot of manual work for data structuring, integration and analysis using spreadsheets and inconsistent file formats. This can be challenging in terms of efficiently performing the data organization that complies with the data structure adapted by other agencies on the same type of data sources. Human resources are limited, and it is challenging to scale the processes in this environment of work practices. Moreover, during the onset of a crisis, in such highly dynamic and uncertain circumstances of crisis response, the data required for integration may not be available to consume in a form that can be readily leveraged to any timely sense-making and decision support. Furthermore, depending on the current phase in the crisis cycle (for the Coordination of Humanitarian Affairs 2014), the data required for any type of analysis may vary largely. Having a framework of data interoperability that allows easy access to existing data sets for integration and analysis can be useful to improve efficiency of the domain workflows with a cost-savvy solution.

Current technologies for structured data storage, integration and analysis are ubiquitous in many business sectors. While the increasing amount of unstructured data sourced from social media during humanitarian events poses new challenges (Purohit et al. 2013), the computing community has been researching smarter ways to transform unstructured data into structured forms. However, among the existing interoperability technologies for structured data , Linked Open Data (Berners-Lee 2009; Bizer, Heath, and Berners-Lee 2009) can still be applied in the humanitarian sector. The Linked Open Data cloud allows sharing the data in interoperable formats, while providing a way for storing knowledge and supporting inference. It has some research challenges, such as the *sameAs* property resolution, however technology can be easily, and efficiently deployed for the domain specific applications in the humanitarian domain, as initiated by the Humanitarian Data Exchange (HDX 2014) project. In the following, we discuss the interoperability, its challenges, solutions, and how it can be adopted in the current context.

Data Graphs, Ontologies, and an Open World Assumption

Given the emergence of Big Data and a desire to make sense of an ever-increasing number of knowledge streams, it is not surprising that the most pervasive technology providers have sought to expand on traditional keyword search methods and ways to represent data as linked sets of statements about entities. Google and Facebook have each introduced versions of Graph Search functionality to their platforms; as Google states, the objective is now less to find answers to direct questions, but rather to "discover answers to questions you never thought to ask". The market for providing services that can traverse semantically-rich directed data graphs is maturing at a very rapid rate, and has achieved a level of utility that is now making its way into the daily lives of a significant number of internet users. It is within this context that we propose technologies to extend current humanitarian organization resources that will provide end users not only with semantically-rich and machine-readable descriptions of vast amounts of data, but also the APIs that will allow them to search these meaningful datasets and retrieve answers to questions they have thought to ask, as well as those they have yet to ask.

An alternative approach to traditional data management strategies represents data using semantic annotation formalism where an annotation is a tuple consisting of the format "subject, property, object", commonly referred to as a triple. Representing the multiple layers of information in triples creates a data model whereby relations between entities are denoted as arcs and labeled with a property name, enabling all of the data to be represented as a graph. The Resource Description Framework (RDF) standard offers a widely-recognized data model for representing triples, ensuring that computing applications can refer to a common language when parsing and consuming these data. The resulting RDF directed data graph can be queried using the SPARQL Protocol and RDF Query Language (SPARQL), a flexible W3C standard for traversing the graph. This strategy results in the same ability to store values associated with records (e.g. the geometry value of a polygon representing a storm track path), but provides a much more flexible structure from which to build out extended capability.

RDF triples and SPARQL provide the foundation for unlocking the inherent meaning of data, cannot by themselves provide all the machine-readable context required for a full understanding of a conceptual domain. In order to provide this next step, metadata and rules describing the "known world" of the domain must be added. An ontology (Gruber and others 1993; Noy, McGuinness, and others 2001) defines this common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain, as well as relations among them. Ontologies provide the building blocks of knowledge discovery, including definitions of vocabulary (lexicons), classes of entities within the domain, properties that describe those classes, and rules for asserting relationships between entities, commonly described by the Web Ontology Language (OWL) standard. Years of research and argument have gone into defining multiple ontologies for several domains, both very abstract (e.g. defining a "Thing") to very focused (e.g. defining the meaning of "nearby" in a spatial relationship context). The byproduct of this work is a widely-shared understanding of the underlying terms and rules comprising an ontology. It is possible, and indeed desirable, to reuse elements from existing ontologies ("bootstrap") in order to create new ones, thus inheriting the community understanding and agreement contained therein.

The final divergence from a traditional data management approach is one of the defining characteristics enabling data graphs and ontologies to realize their potential: the Open World Assumption (OWA). With the OWA, if something is not known to be true about a domain, it is not assumed to be false; rather, it is assumed to be unknown. This is in contrast to a Closed World Assumption (CWA), whereby what is known not to be true is assumed to be false. In an operational context, the CWA is often manifest in a Relational Database Management System (RDBMS) approach: anything not explicitly contained in a database schema is not considered relevant to a domain, and is ignored. This has practical implications that range from a user's ability to access needed data to the administrator's ability to extend or change a rigid database schema. The structure of RDF and ontologies lend themselves to the OWA, providing the flexibility necessary to intelligently extend the data graph when new knowledge is introduced.

Linked Data Principles

The OWA described above offers not only a path to creating new knowledge, but also the ability to provide interfaces to knowledge existing within an ecosystem of operators adhering to this worldview. Ontologies provide a self-description of the domain as a whole, but that by itself does not provide all the necessary information for a completely interconnected data web. Still necessary are self-describing data elements that refer to canonical representations of real-world objects, so the descriptions of the connections between distributed datasets encoded in RDF and ontologies are referring to the same subject.

Linked Data is about using the web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods. Linked Data builds on several existing web enablers, such as HyperText Transfer Protocol (HTTP) and Uniform Resource Identifiers (URIs), in order to provide a standard method for representing objects or concepts. URIs serve as global identifiers that "name" things, and via HTTP requests, information about the things can be returned to users in various formats, including RDF and HyperText Markup Language (HTML). For example, the URI `http://www.geonames.org/4140963/` refers to Washington DC, the capital of a political entity, the United States, itself represented by the URI `http://sws.geonames.org/6252001/`, in the GeoNames gazetteer of place names. These two URIs represent two real-world administrative entities, and can be combined using a RDF property "hasCapital". Due to the fact that GeoNames is a trusted source of these representations, users have reasonable assurance that if they use these URIs in their own Linked Data application, they will be referring to the same thing.

Using this approach, humanitarian entities can similarly position themselves as URI providers of entities related to their domains of expertise, and further enable a network of actors to understand each other. A further step toward useful interoperability, beyond agreement on the meaning of information, is the trust that accompanies authoritative information. In the Semantic Web stack this is something of a final step for reaching the full potential of interoperability. Knowing that the providers of Linked Data are authorities on the canonical subjects they describe breeds confidence among consumers of distributed, remotely-hosted knowledge. Several mechanisms and protocols exist for authenticating sources of information, such as digital certificates and signatures that can be verified using ubiquitous libraries and applications. Further trust can be established with enhanced, updated metadata, so that end users not only know the source, but also the validity of sometimes rapidly-evolving situational information. Also, not every humanitarian operational unit is required to adopt Semantic Web based technology but rather the main centers of complex information management, and the field units can still operate with traditional means (e.g., excel sheets) of data exchange to the main centers, where apps or plugins can seamlessly integrate the exchanged data to main centers in interoperable forms.

Geospatial-Semantic Reasoning for Enhancement of the Humanitarian Knowledge Base

With Linked Data represented as RDF triples, and operating within the rules of a domain ontology, a natural opportunity exists for extending implicit knowledge through an evaluation of the explicit relationships already represented. This can be achieved by utilizing semantic reasoners that can evaluate both semantic and geospatial properties of data, enabling applications to convey not only explicit spatial and temporal linkages, but also infer any implicit relationship involving the logical expressions of subject and property values in the multi-dimensional semantic space.

The potential benefit of pursuing such an approach is best presented in a practical example. To specifically answer the question: Where are the vulnerable populations in the path of Hurricane Boris? Here, semantic representations (including geospatial property values composed as Well Known Text (WKT) strings) of humanitarian data can be evaluated along with social and demographic knowledge bases containing demographic data in order to infer triples that represent the answer to the question. Specifically, considering the following conceptual example triples:

```
<Hurricane_Boris> <hasImpactAreaGeometry>
    <[WKT polygon of hurricane
    track and impact areas]> .
<Fairfax_VA> <hasGeometry>
    <[WKT polygon of county
    boundaries]> .
<Fairfax_VA> <hasPopulation> <Pop_A> .
<Pop_A> <hasResiliencyMeasure> ``Low'' .
```

We can infer, through a combination of transitivity (as specified in OWL) and geospatial intersection rules, that:

```
<Hurricane_Boris><affects><Pop_A>
```

This triple did not exist until we performed logically-consistent inference functions on the original set of triples. Using these techniques, triples in two or more datasets were combined through first-order logic to deduce new facts from stated ones. This simple example can be further enhanced with 1) inference logic applied to geospatial relationships, such as those defined in the Regional Connection Calculus (RCC) (Cohn et al. 1997; Hart and Dolbear 2013) or the Dimensionally-Extended 9 Intersection Model (DE-9IM) (Clementini, Sharma, and Egenhofer 1994) additional OWL properties lending themselves to inference, such as transitivity, symmetry, functionality, or inversibility of relationships within the data. Using these capabilities, humanitarian end users will be empowered with the ability to understand much more than is explicitly asked of the data, and can carry out the above example in the context of operations, rather than focusing on asking a series of rigid questions.

The Humanitarian Exchange Language (HXL) Prototype

The initial prototype version of the Humanitarian eXchange Language (Keßler and Hendrix 2015; Keßler, Hendrix, and Limbu 2013), developed at the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA), was entirely built using Semantic Web technology and the idea of Linked Data (Berners-Lee 2009; Bizer, Heath, and Berners-Lee 2009). This approach seemed appropriate for a number of reasons:

- The data stays with the owner. There is no central repository or super system, nor a standard that would have to be imposed on the partners to implement. Instead, everyone can stick to their data management systems and only expose the data for exchange and interlinking.
- The vocabulary used for data annotation is extensible. Coming up with a fixed schema to cover the very broad range of humanitarian data is hopeless, so providing a core vocabulary that can be extended as needed is much more realistic.
- For many sub domains that humanitarian data relies upon, existing and well-established standards can be reused. Examples include geospatial (Open Geospatial Consortium 2012), social (Brickley and Miller 2010) and metadata (Dublin Core Metadata Initiative 2012).
- Standards for data exchange formats, a standardized query language, and storage systems implementing these are in place and have been demonstrated to scale. Hence, there is no need to design, implement, test, and maintain a new, HXL-specific API.
- Leveraging external data sources that are not per se humanitarian, but still useful to understand the situation on

the ground, is straightforward and, in fact, one of the design principles of Linked Data.

In order to import existing humanitarian datasets, which have been (and still are) exchanged mostly in more or less standardized spreadsheets, a prototype import tool has been developed that makes heavy use of the reasoning capabilities to guide the user through the process of semantically annotating a spreadsheet. The HXLator[1] is a web application that allows users to upload a spreadsheet and then subsequently convert its' contents to RDF. It starts out by asking the user what kind of objects the spreadsheet has information about, and then uses the HXL vocabulary's (Keßler and Hendrix 2015) domain and range definitions for properties to determine what properties the spreadsheet contents can be mapped to. The user is guided through the process of mapping every field in a single data row, and when done, specifying the range of data rows in the spreadsheet. Once the data has been converted to RDF, it is uploaded to an private incubator triple store for review, before it gets approved and ultimately pushed to the public triple store. Throughout this process, the HXLator makes heavy use of SPARQL and SPARQL update queries to interact with the triple stores. The translator created in such a session can be saved and reused for other spreadsheets that use the same layout.

If we take a step back from data formats and query languages, the central element to make the different nodes in our humanitarian data network interoperability is a common language. This common language covers the core terminology used in the domain, and a clear definition of its meaning. If this terminology – whether we call it a vocabulary, a taxonomy, or an ontology – is done right, data formats and storage methods become a secondary issue, since any column in a spreadsheet, table in a database, element in an XML file, or resource on the web can be described in terms of the agreed-upon terminology. This terminology should therefore be the central concern of HXL phase II. As in the prototype version of HXL (Keßler and Hendrix 2015), existing standards such as schema.org or PROV (Moreau and Missier 2013) should be embraced. Furthermore, the HXL vocabulary can be extended to go beyond data management, such as to assist humanitarian response coordination functions of resource seeking and supplying for applications (Purohit et al. 2014).

Next Steps

The authors of this paper argue that many of the foundational building blocks that would enable a broad network of humanitarian Linked Data publishers exist and are quite mature. From a technology procurement perspective, these options are more attractive than at any time in the past.

The Semantic Web capabilities outlined above are all very much rooted in ubiquitous internet technology and protocols. Regardless of the usefulness from an operational information sharing perspective, convincing organizations with funding constraints procuring these resources has traditionally been a difficult task. However, within the past decade, the rise of the number of large internet players offering Infrastructure as a Service (IaaS) – more commonly known as commercial cloud computing – has lowered the cost barrier to entry and enabled governments, large companies, and startups to deploy their entire suites of IT assets virtually. As more providers have emerged, competition has led to dramatic price decreases in hosting websites, data stores, and other such computing applications. From a Linked Data perspective, this has the potential to include ever more organizations into the web of data, and provide humanitarian actors with more resources for their operational use.

Several alternatives exist for deploying IaaS. Google Compute Engine, Microsoft Cloud, IBM SoftLayer, and Amazon Web Services (AWS) offer users the ability to create virtual environments that have significant applicability to the humanitarian domain. These resources offer application developers the ability to create highly available, fault-tolerant capabilities that can withstand spikes in user demand (e.g. during a rapid-onset disaster when information is being requested at a high rate), as well as the inevitable failure of hardware and software. For example, AWS has several service offerings geared toward ensuring uptime in applications. Their data centers are organized into Regions across the world, and further subset into Availability Zones, allowing users to deploy applications in a distributed manner for providing failover assurance. Disaster recovery, in a computing sense, is easily achieved through distributed data store and application code backups (using AWS's Simple Storage Service, or S3, for example). In addition, using AWS AutoScaling, administrators can set conditions to where new server resources can be instantly spawned and added to a functioning cluster when demand increases, and can scale down when that demand subsides. Further efficiency of delivery can be achieved by leveraging a Content Delivery Network of edge locations in various parts of the world to cache and deliver data with reduced latency. With these options, publishing Linked Data is arguably more cost-efficient and feasible than in the past.

Conclusion

We presented a case for how structured data technologies, such as the Linked Data originating from the Semantic Web research area, can be applied to address the challenge of information interoperability in the humanitarian domain. We observed the need for work on the transformation from unstructured to structured data, as well as iterative development of technological solutions and a formal domain specification via data schemas and ontologies. Using the Humanitarian Exchange Language (HXL) ontology as an example, we described the feasibility and challenges of adopting the structured data technologies in the humanitarian domain. We conclude that domain experts also need to come forward to help design the foundation of interoperability, the ontology to describe the domain to improve the efficiency of data management and integration.

[1] See https://github.com/hxl-team/HXLator.

References

Berners-Lee, T. 2009. Linked Data – Design Issues. Available from http://www.w3.org/DesignIssues/LinkedData.html. Accessed Oct 30 2014.

Bizer, C.; Heath, T.; and Berners-Lee, T. 2009. Linked data-the story so far. *Int. J. Semantic Web Inf. Syst.* 5(3):1–22.

Brickley, D., and Miller, L. 2010. FOAF Vocabulary Specification 0.98. Available from http://xmlns.com/foaf/spec/20100809.html. Accessed Oct 30 2014.

Clementini, E.; Sharma, J.; and Egenhofer, M. J. 1994. Modelling topological spatial relations: Strategies for query processing. *Computers & graphics* 18(6):815–822.

Cohn, A. G.; Bennett, B.; Gooday, J.; and Gotts, N. M. 1997. Qualitative spatial representation and reasoning with the region connection calculus. *GeoInformatica* 1(3):275–316.

Dublin Core Metadata Initiative. 2012. DCMI Metadata Terms. Avaliable from http://dublincore.org/documents/dcmi-terms/. Accessed Oct 30 2014.

EM-DAT. 2014. Natural Disasters Trends. Available at http://www.emdat.be/natural-disasters-trends. Accessed Oct 30 2014.

for the Coordination of Humanitarian Affairs, U. N. O. 2014. UNOCHA Cluster Coordination. Available at http://www.unocha.org/what-we-do/coordination-tools/cluster-coordination. Accessed Oct 30 2014.

Gruber, T., et al. 1993. A translation approach to portable ontology specifications. *Knowledge acquisition* 5:199–199.

Hart, G., and Dolbear, C. 2013. *Linked data: A geographic perspective*. CRC Press.

HDX. 2014. Humanitarian Data Exchange (HDX) project. Available from http://docs.hdx.rwlabs.org/. Accessed Oct 30 2014.

Keßler, C., and Hendrix, C. 2015. The Humanitarian eXchange Language: Coordinating Disaster Response with Semantic Web Technologies. *Semantic Web Journal* 6(1):5–21.

Keßler, C.; Hendrix, C.; and Limbu, M. 2013. Humanitarian eXchange Language (HXL) Situation and Response Standard. Available from http://hxl.humanitarianresponse.info/ns. Accessed Oct 30 2014.

Moreau, L., and Missier, P. 2013. PROV-DM: The PROV Data Model. W3C Recommendation, available from http://www.w3.org/TR/2013/REC-prov-dm-20130430/. Accessed Oct 30 2014.

Noy, N. F.; McGuinness, D. L.; et al. 2001. Ontology development 101: A guide to creating your first ontology.

Open Geospatial Consortium. 2012. OGC GeoSPARQL – A Geographic Query Language for RDF Data. Available from http://opengis.org/standards/geosparql. Accessed Oct 30 2014.

Purohit, H.; Castillo, C.; Diaz, F.; Sheth, A.; and Meier, P. 2013. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday* 19(1).

Purohit, H.; Hampton, A.; Bhatt, S.; Shalin, V. L.; Sheth, A. P.; and Flach, J. M. 2014. Identifying seekers and suppliers in social media communities to support crisis coordination. *Computer Supported Cooperative Work (CSCW)* 23(4-6):513–545.

ReliefWeb. 2013. Annual Disaster Statistical Review 2012: The numbers and trends. Available from http://reliefweb.int/report/world/annual-disaster-statistical-review-2012-numbers-and-trends. Accessed Oct 30, 2014.

UNESCAP. 2013. Statistical Yearbook for Asia and the Pacific. Available at: http://www.unescap.org/stat/data/syb2013/F.5-Natural-disasters.asp. Accessed Oct 30 2014.