Issues in Incrementally Adding Better Semantics to Knowledge Graphs

Gary Berg-Cross

RDA/US Advisory Group

Abstract

Knowledge graphs (KGs) employ a wide range of semantic resources. However, as is true of complex information systems, harmonizing rich semantic resources requires effort and involves trade-offs. There are practical reasons to start with modest semantics, and then incrementally add enhanced semantic improvements. For this process there are a number of active research projects that are developing light, incremental approaches, methods and tools to support an expanding semantic KG space that has addressed semantic alignment and harmonization. These projects include methods for using existing semantic relations and entities harmonized across controlled vocabularies, glossaries of definitions and ontologies. This article discusses examples of incrementally improving the semantics of less formal schemas that over time is helping to semantically unify richly interconnected heterogeneous data using newly adopted and agreed upon methods.

Introduction

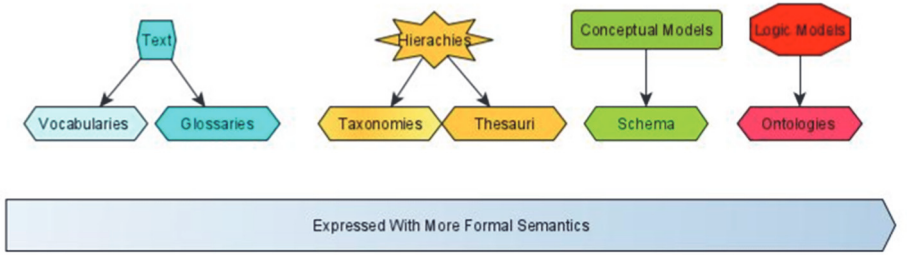
IT IS COMMON PRACTICE to engineer complex information systems using iterative processes to provide incremental improvements (Darrin and Devereux, 2017). This is reflected in principles of extreme programming (XP) which is iterative and incremental and tries to address only a few modeling issues during each phase of work (Choudhari and Suman, 2010; & Dalalah, 2014). A form of extreme programming called the eXtremeDesign (eXD) approach has been used for the development of semantic resources such as knowledge graphs (KGs), ontology design patterns (ODPs), and ontologies. Indeed the start small/iteratively improve strategy of making incremental improvements, starting from some simple semantics, was an implied part of the Semantic Web (SW) vision. It can be seen in the semantic spectrum diagram (Figure 1) that starts with terms and glossaries that have textual definitions and moves to a series of increasingly semantically precise and expressive definitions of conceptual entities that at the far end of the spectrum are represented in some formal language. This original SW vision was that over time more data will be represented in forms processable by computing machines to enable efficient and effective

semantic exploration of the Web by bringing "structure to the meaningful content of webpages, creating an environment where software agents can roam across webpages in order to carry out sophisticated tasks for users" (Hogan, 2020). In this vision incremental improvement choices should start at the low end of the spectrum of semantic resources, which might be a single word or phrase and its conceptual definition. At the high end we arrive at a very expressive formal ontology with structured and unambiguous ways of representing domain information. Formal languages used to represent information are intended to specify processable relationships between formally conceptualized data elements and be part of the SW vision, and to employ precise URIs for relationships and properties. Implied in the semantic spectrum model are "bottom-up" engineering approaches leveraging more informal information. This can work, as shown by KG efforts starting with relatively unstructured descriptions of a domain. We can encode initial concepts from words and their implied relations informally understood by domain experts. These concepts can then be simply structured into a prototype model that identifies key concepts. From a simple start increased semantic specificity removes ambiguity and affords a growing degree of sharing and interoperability. The idea of incremental development is attractive since it is a way to use the growing abundance of information and semantic resources hosted on the web and available for applications.

However, in light of the experience with a variety of semantic resources since the original semantic web formulation, there are more than simple, pointwise improvements to be made formalizing web data by say expressing some definition in OWL. In addition to this there are improvements to such related issues as coverage of relevant information, factual or timely correctness, overall structures between items, and methodological questions of how any of these changes may be done efficiently and effectively. To address this I provide examples of both simple, somewhat isolated semantic enhancements, and ones that are broader. I also provide examples of some systematic methods, such as harmonization across a suite of concepts. A challenge of an incremental approach is that while simple concept definitions can serve as links between things on the semantic spectrum, there may be numerous ways of defining a concept. The word "forest" is used conversationally, but there is no universally recognized precise definition. More than 800 definitions of forest are

recognizable around the world, including off center relations to forest nurseries and forest roads. To incrementally improve semantics as one moves along the spectrum core relations such type and part-whole relations can be used to reduce ambiguities in a conceptual space. Consistent with this idea early ontological engineering practices aimed at reaching the formal end of the semantic spectrum emphasized that ontology development is necessarily an iterative process with regular revisions, debugging, and progressive deepening. (Noy and McGuinness. 2001). This general process is equally true for KG development.

Figure 1: Semantic Spectrum from Defined Terms to Formally Defined Concepts based on (MA, 2021)



Data abundance provides opportunistic construction of products like KGs by leveraging the large amount of available data, including semantic resources along the semantic spectrum. However, real-world KGs are complex. The data, while vast, includes informal types, is usually incomplete, and is not harmonized. Thus, early phases of work with raw data do not easily reflect a full reality-based model. There are risks to assume that one can develop a fully validated domain-spanning semantic model all at once. As noted by Elsaleh *et al*. (2019), "semantics add further overhead to data delivery, and the processing time to annotate the data with the model can increase the latency and complexity in publishing and querying the annotated data."

For these reasons and others, in practice, most KG projects are largely data driven from the bottom and do not build or use a full and semantically rich, domain ontology. Instead, projects search for low-hanging fruit and tractable semantic resources, such as a centralized, crowd-sourced approach using Wikidata as the foundation or using generalized vocabularies. Still another approach is to leverage a slimmed down ontology (Heller *et al*., 2018). Work often starts with what are called

lightweight ontologies (Giunchiglia and Zaihrayeu, 2007). These aid developers (and users) by allowing relatively fast annotation of data with information from glossaries. These are midway along the expressiveness dimension of the semantic spectrum. Lightweight ontologies and related models are largely descriptive and include groups of concepts, concept taxonomies showing sub-classes, and simple, conceptual relationships between concepts. Lightweight ontologies have only a modest number of axioms and relations, and may focus on taxonomic and structuring relations. Heavier or richer ontologies include axioms and constraints beyond hierarchical ones to clarify the intended meaning of the terms involved in a domain. There are a variety of trade-offs between light and heavy models. Obviously, richer axioms done well are closer to domain reality. But a benefit justifying a lightweight approach is to save query processing time when complex relations are involved.

A big advantage of starting with low-hanging resources and lightweight semantics is the desire to make rapid progress. So we typically see that KGs may be implemented using data with limited semantics, such as the use of RDF triples or property graphs with no schema. As a result projects may have masses of data with no consensus on core semantics or what models are reflected in the data. Even in a structured form like RDF, data may be developed from different vocabularies and different perspectives on the data and largely be stored in "dispersed forms in a number of autonomous information silos" (Guizzardi, 2020). As Heflin and James Hendler (2000) put the resulting challenge of integration: "To achieve semantic interoperability systems must be able to exchange data in such a way that the precise meaning of the data is readily accessible and the data itself can be translated by any system into a form that it understands." We need some degree of deep knowledge to support domain reasoning to fulfill this SW vision.

FAIR Guiding Principles

A step towards sound and incremental data management practices regarding new knowledge generation and discovery by individuals and organizations was taken in 2016, with the publication of the 'Findability, Accessibility, Interoperability, and Reuse' (FAIR) Guiding Principles for scientific data management and stewardship (Wilkinson *et al*., 2016). The FAIR principles include enhanced semantics for machine-actionability (*i.e*.,

the capacity of computational systems to find, access, interoperate, and reuse data). The "I" in FAIR is concerned with putting machine-readable knowledge on the web, and incremental semantic technology helps to achieve this by developing comprehensible structure and context that makes data easier to reuse and integrate with other data.

The Internet of Things as an example

The Internet of Things (IoT) is one area of interest to KG development that does not yet have a consensus on top-down models or ontologies. This reflects the more general fact that unitary efforts to develop top-level ontologies or even broad and deep domain ontologies have had problems (De, Suparna, Zhou, and Moessner, 2017). Because of the breadth of entities involved in IoT, formal agreements on semantics tend to be avoided to make KG construction quicker. Instead, as noted before, a bottom-up process is often the initial guide. This leverages analysis of implied meanings, as understood by domain experts and/or data analysts, found in structured and linked data. This makes sense since the choice of semantic simplicity avoids complex metadata documentation and encourages faster adoption by end users who can easily grasp the concepts. However, ignoring top-down semantics can sacrifice system quality, making it difficult to interpret query results against intended or reliable agreed upon meaning. For example, without an effective naming authority for Web data, it can happen that different KGs refer to the same thing by different names, and the same names may have quite different meanings (Alexopoulos, 2020). This problem is well known from earlier work on conceptual models of data (Hull, 1996). A modeling example is the difference between a glossary for cars versus the more general one of automotive vehicles, which would contain trucks. Manually resolving these differences can be challenging and involve many trade-offs. Creating a semantic data model is a labor-intensive process, and requires a sound understanding of the selected domains within a KG's scope along with the relevant ontologies.

Data Heterogeneity Tradeoffs and Challenges

For many reasons, KG building tends to fall back on some idea of incrementally adding semantics as part of later stages of iterative development. However, there are several challenges for improving KG

semantics, such as data heterogeneity. Barriers exist to continued, incremental progress, such as finding an easy, effective way to create richer vocabularies, remove ambiguity, and integrate richer semantics into schemas with appropriate constraints and relations. The struggle starts with the problem of conceptual heterogeneity, such as contradictory structures and/or levels in different taxonomies or other lightweight semantic resources. Groups crafting definitions (or building other semantic resources like conceptual models) have varying experience and ability and use different methods. This situation can innocently result in unintended heterogeneity. One may use some knowledge engineering to craft semantically harmonized definitions, but this can be time-consuming, error-prone, and a tedious process. Domain experts can quickly lose interest in such work. Indeed, different domain experts may use varied implicit background knowledge to understand, and later define, concepts with the same name. Experts and knowledge engineers may locate the same targeted concept differently within some conceptual space or hierarchy. In turn, the use of significantly different distinguishing concepts creates conceptualization mismatch in a KG. As a result, the KG may misalign data based on their different underlying source models, even when attributed to hardworking domain experts.

Incremental improvements may also be challenged by the trade-off between coverage and correctness. Coverage is concerned with whether the KG has all the required or desired information. Effectively the answer is always no, in the sense that domain knowledge is indefinitely extensible, and development teams are motivated to look for new ways to provide value to domain users. And it is also true that new sources of data, information and organizational schemes, like language, emerge over time. However, as the coverage increases, the likelihood of a conflict or contradiction also increases. Approaches to the trade-off between coverage and correctness may be addressed differently in particular KGs when difference in the coverage occur and new data sources are introduced. KGs may also need to extend their semantics to describe different realms or regions of the world at the same level of detail and/or from a different perspective. Similarly, differences in granularity occur when we have the same perspective, but at different levels of detail. This occurs, for example, with geographic maps of different scales. Difference in perspective (difference in scope) occurs when two data sources describe the same region of the world, at the same

level of detail, but from different perspectives (*e.g.*, a political map vs. a geological map). For all these reasons revisions and modifications in a KG lifecycle may result in ambiguity, redundancy, and modeling inconsistencies that need to be addressed over time.

In general, achieving a base level of interoperable results given the heterogeneity among different information systems is difficult (Maciel, 2017). More than a simple change of representation is needed as part of data integration and harmonization. The simple fact is that in order to achieve a useful degree of interoperability between datasets, either the datasets need to use the same (set of) ontologies, or the ontologies need to be aligned and mapped. How to develop efficient alignment and mapping is one of the areas of improvement emerging from KG research.

Examples of Incremental Improvements

To give the flavor of work that addresses some of the challenges KG and related efforts face, six types of incremental semantics are briefly illustrated, starting with the simple case of improving and harmonizing identifier systems. This seems a relatively simple problem to address, but still exists and is a major concern of the FAIR principles. We then consider four specific ontologies. This section ends with a discussion of automated techniques.

Clarifying Identifier Systems

The wide range of semantic resources often use different identifier systems. For example, the W3C Organization Ontology (ORG) expresses information about organizations, *i.e.*, companies and institutions, including governmental organizations. The focus is on organizational structure (*e.g.*, sub-organizations and classification of these), along with reporting structures (roles) and facility locations (Reynolds, 2014). In contrast, the e-Government Core Vocabularies that were developed in order to provide a minimum level of semantic interoperability for e-Government systems use a different system that covers overlapping concepts with a focus on public services, public organizations, and public services (European Union, 2015; Gerontas. 2020). There are many other examples, and so developing a unified semantic expression of identified semantic resources becomes a more difficult task as more resources are assembled into a KG. In the business domain, the euBusinessGraph ontology represents an example of

using a rather specific lightweight semantic model approach to standardize identifier systems within their area of interest. The euBusinessGraph ontology starts by systematically combining and reusing termed concepts from existing ontologies, such as the previously mentioned EU Core Vocabs: W3C Org, as well as others (Roman, 2021). The result is a revised model integrating several semantic resources, such as vocabularies, into a more expressive and detailed model that includes extensions to "application profile, RDF Shapes and data provider mapping documentation." (Roman *et al*., 2021).

Extensions and Improvements to FOAF

Refinement and extensions of semantic resources to handle expanded requirements are among the typical increments needed to integrate semantic resources. An example of incremental improvement can be seen in the movement from the early conceptualization of Friend of a Friend (FOAF) (Brickley & Miller, 2018). FOAF is a widely used lightweight social network "core vocabulary", but it has vocabulary areas with little or no adequate semantic coverage. This combination becomes a problem as the vocabulary keeps being reused without improvement. The answer is to avoid slavish reuse and to make incremental extensions. Examples of this makes the FOAF case illustrative of incremental development. There are now numerous refinements and extensions for particular areas. This process may start with expansions of the subclasses and with deepening the classes found in early versions of FOAF. Refinements may take the form of defining inclusions between classes and relations, and/or of refining features and restrictions of the relations. A vocabulary example is the extension of the definition of "landform" as "a feature on the Earth's surface that is part of the terrain." to including formative processes (*e.g*., volcanic process or tectonic movements) or constituents such as magma as part of a volcano system.

A useful example of a social network extension to meet expanded requirements comes from considering new virtual, social relations (El Kassiri and Belouadha, 2017). These include worked out examples covering extensions to social networks in a variety of behavioral-and health-related research areas. These capture more of the rich interactions of social networks (Amit et al., 2020). Enriching FOAF in this way uses the best practice of building on the existing (FOAF) model and illustrates the

practice of incorporating ideas from other ontologies. To do this one can reuse one or more lightweight ontologies, adding extra properties and classes as needed. In terms of ontological engineering practice, this reflects adding new competency questions (*i.e.*, the questions a knowledge base can answer) which are not addressed in a current version of a KG or its underlying ontology.

Another FOAF extension example uses ontological imports from the Food and Agriculture Organization's geopolitical ontology (Kim and Viollier 2013). The result is a richer artifact, called FOAF+. It can be used to describe new types of social ties, interactions and new entity features or attributes not included in the earlier, base and lightweight FOAF. Another aspect of incremental improvement is the quality of conceptual, not just representational, expressiveness. Unlike the early idea of the semantic spectrum, one is not just increasing the expressive language around a concept; one has a quality meaning that is prior to the expression of the knowledge. Otherwise, we face, adapting an old expression, the situation of "imperfect modeling information in; imperfect understanding out."

Schema.org and Bioschema

Annotating data for KGs using a lightweight, standard schema is another situation that illustrates semantic enhancement. Schemas represent a target for mapping and are one way to support improved alignments. An example comes from the use of a master data hub like the Schema.org markup vocabulary. Schema.org reflects an effort to standardize lightweight, annotating vocabularies to reduce data heterogeneity, as well as to ensure that websites are more uniformly indexable for search engines and other web services (Guha *et al*., 2016). As previously discussed, KGs typically start with the low-hanging fruit available from annotations that can be extracted from websites. An incremental improvement is to standardize these by verifying them against the relevant domain part of the Schema.org vocabulary.

Extensions of Schema.org for the bio-science realm illustrate some of the progress. For example, extensions to Schema.org as part of Bioschema work (Franck. 2018) allows searching for and finding data about specific biological entities (*e.g*., particular genes, proteins, and taxa). This enhancement uses entity profiles. An example of this is the idea of TaxonName. This is used to specifically annotate taxonomic name registries.

Guidelines for use are provided that describe how to leverage existing vocabularies such as Darwin Core or Wikidata. Bioschema can be used to improve biology oriented KGs by enabling semantic cross-linking between any KG that extracts data from Bioschema marked-up websites.

The Bioschema enhancement also follows FAIR principles. It includes semantics to help discover data repositories storing experimental results, along with the storage location of specific biological samples. To support this, relevant controlled vocabulary terms drawn from existing ontologies have been imported. A cited example is the protein profile which requires a unique identifier, and recommends listing transcribed genes and associated diseases. For proteins Bioschema points to recommended terms from the Protein Ontology and Semantic Science Integrated Ontology.

As an incremental improvement to Schema.org, Bioschema specifications go beyond simply adding new types and properties for biological entities. It includes more structure, providing constraints on the Schema.org model. These constraints capture and require things like minimal information properties agreed by the bio-community. These fall into categories of mandatory (M), recommended (R), or optional (O). Another enhancement is that the cardinality/occurrences of properties have been added.

Success with semantic enrichment such as Schema.org generates support from tools to automate some of the activity. As an example, a semantic validator has been developed to help ensure the syntactic correctness and completeness of the annotations from a Schema.org perspective (Panasiuk *et al*., 2019).

Sensor, Observation, Sample, and Actuator ontology (SOSA) an Example of a Design Pattern Approach

SOSA is a lightweight but a general-purpose pattern-based specification for modeling the interaction between the entities involved in the acts of observation, actuation (actions triggered by observations), and sampling. As an incremental example, SOSA is the result of rethinking the W3C-XG Semantic Sensor Network (SSN) ontology. It reflects changes in scope and target audience, technical developments, as well as lessons learned over the past years. SOSA, as well as its base SSN, are examples of an incremental approach to semantics using ontology design patterns (ODPs)

(Gangemi, 2005). ODPs help address the reuse problem of complex semantic resources, such as an ontology like DOLCE, where only certain “useful pieces" of a comprehensive (foundational) ontology, may be of interest to a KG. This is based on the observation that the cost of reuse from a large, but shallow ontology, may be a higher cost on resources than developing from scratch a scoped ontology for particular purposes. ODPs can serve as a step towards more structured and semantically rich metadata, even to extend something like Schema.org markups. ODPs reflect the understanding that often to practically solve semantic problems, it is productive to agree on minimal requirements imposed on a relevant family of concepts (Kuhn, W. 2009). ODPs (aka microtheories) represent small, well engineered, coherent, minimally constrained schemas. Light, general ontology patterns function as a modular consistent core that can serve as a starter set from which more varied and detailed models or, as needed, larger ontologies can be built. In effect ODPs serve as an initial constraining network of “concepts” within a common framework using vocabularies allowing people to incrementally extend and align them for various purposes. The overall impact to support some degree of common interoperability, including easy data sharing via a KG using one or more ODPs.

The eXD methodology, mentioned earlier, uses an agile approach to ontology engineering and focuses on the reuse of ODPs (Presutti *et al*., 2009). Among ODP best practice qualities that enable incremental semantics are explicit documentation of design rationales, and the use of best re-engineering practices to facilitate reuse. It is also worth noting that a KG design around a family of ODPs can provide a very concise but informative view of the overall content of a KG (Asprino *et al*., 2021).

ODPs like SOSA support a progression from the light semantics of something like FOAF to heavier semantics that helps avoid legacy and silo building problems (Janowicz *et al*., 2019). As an example of how ODPs afford easy improvement, SOSA uses axiomatization but does not make any formal restrictions on how to use observable properties. This allows for alternative observational models. Observed features of interest, for example, are not restricted to objects like roads. They also can include events such as rush hour traffic. As with other semantic products, over time there has been a recognition for refinement in ODPs. In the case of SSN, which evolved into SOSA, this process included recognizing the need for a central concept

of observational sampling over time. Another recognized improvement was that observations may not be carried out on the entire feature, but on samples of a feature and/or as part of sensing some spatiotemporal region that serves as a proxy for a feature (Taylor *et al*., 2019).

Enhancing Earth Sciences Ontolgoies: adding EnvO axioms to SWEET

A fifth illustration of incremental semantic improvement comes from experiences with broad ontologies like the Semantic Web for Earth and Environmental Terminology (SWEET). SWEET is a lightweight ontology with broad coverage, but sporadic definitions that historically served as a starting point for concepts within the Earth Sciences (DiGiuseppe, Pouchard, and Noy. 2014). Often, richer semantics were added for particular domains. A more semantically richer, but topically overlapping, ontology is EnvO, the Environmental Ontology. EnvO includes semantically controlled descriptions of environmental entities and rich axioms. It thus serves as a quality semantic resource for research and is widely cited. For example, the Darwin Core glossary uses EnvO for habitat descriptions. Over its life, EnvO has been continually extended beyond its initial goal to represent biomes, environmental features, and environmental materials pertinent to genomic and microbiome-related investigations. The need for environmental semantics is common to a multitude of fields, and thus EnvO's use has steadily grown since its initial description. Its scope has expanded, been enhanced, and generalized, so the ontology can support its increasingly diverse applications, as shown by the range of updates in a recent release (EnvO, 2021). One notable example of a recent extension is as a semantic resource for Cryosphere concepts (Berg-Cross and Vardemann, 2020). Work on a common Cryosphere model to be added to EnvO actually started with a 14-hour hackathon on glaciers (Glacier Hackathon, 2019). The session leveraged and harmonized a focused portion of a rich collection of definitions developed by the World Meteorological Organization’s (WMO) Global Cryosphere Watch (GCW). This GCW work expanded on prior work on sea ice ontologies (Duerr *et al*., 2015) and had recently collected some 27 cryospheric glossaries containing a total of 4147 terms. Importantly, semantic analysis showed that only 2249 were unique and terms could be organized into useful categories; namely, those

- that were well-formed and documented and not problematic from a semantic standpoint,
- where multiple definitions could be coalesced into a single definition, and
- where the terminology was inconsistent and therefore problematic from a semantic standpoint, and where community resolution was needed to either agree on a definition or to split the terms up into separate entities, *etc*.

The subsequent hackathon's objective was to develop a refined conceptual model organizing relevant terms into glacial object types, features, composition (*e.g*., frozen water matter) and processes. The overall hackathon experience of building a conceptual model with some ODP structures was successful enough for participants to seek a way to continue this work on a regular basis. A goal was to use this as a way of aligning and enhancing the SWEET and EnvO ontologies. This was an important test case because both ontologies were independently developed, but both ontologies contain many of the same important semantic resources of the environmental and earth science (ESS) domain. Discussions had been underway about how to align portions of them. A domain like the cryosphere with some harmonization efforts completed presented a good test area for ontology alignment and enrichment. To support this, a Semantics Harmonization cluster was formed with the Earth Science Information Partners (ESIP) Semantic Technology Group to develop a harmonized cryospheric glossary leveraging cryospheric terms in both Envo and SWEET.

As a start, GCW's harmonized definitions were used to add content to the EnvO while simultaneously mapping the results to existing classes in the SWEET Ontology. EnvO's ontology is more richly axiomatized than SWEET since it is part of the Open Biological and Biomedical Ontologies (OBO) Library, and employs recommended practices and technologies for developing expressive and interoperable ontologies in OWL. As ontologies like EnvO and SWEET are enhanced, they in turn can support their domains for KG development.

Below is an example of a semantic update to the EnvO description and axioms for the concept of "ice shelf":

- An ice shelf is an ice mass attached to the coast

- An ice shelf is at least 2 meters in thickness
- An ice shelf forms where a glacier or ice mass flows down to a coastline and onto the ocean surface and
- An ice shelf grows by annual snow accumulation or by the seaward extension of land glaciers.

Some corresponding Envo Axioms are:

- partially surrounded by some atmosphere
- attached to some sea coast
- has quality some buoyancy
- adjacent to some marine water body
- formed as result of some snowfall
- a land ice mass
- formed as result of some mass ice flow

As a whole, SWEET remains less axiomatized than EnvO, but has been updated to use the same harmonized definition that ENVO does. And to support ontology alignment, axioms have been added to SWEET asserting such things as a "closeMatch to EnvO Ice shelf". This allows cross-fertilization with SWEET's long-standing usage in the Earth and environment domain, and offers a pathway for its incremental development.

Relations like "closeMatch" expressed in RDF are part of the lightweight SKOS standard. They provide only modest expressivity for mappings such as mentioned above to relate terms in EnvO and SWEET. A popular formalism for relating terms is SKOS, which uses RDF to provide some formalization of various types of controlled vocabulary. These include classification schemes, subject heading lists, and taxonomies. This promises some degree of automation for finding relevant terms and for aligning similar terms. But using SKOS to define vocabularies and term relations can lead to problems down the road. Since it has semantic limitations, some of its modeling is suggestive, rather than constraining enough to reduce ambiguity. For example, SKOS doesn't provide axioms to explain the similarity and work can struggle to upgrade the SKOS model into a more expressive language like OWL. There is no straightforward path between the two languages without conceptual analysis (Jupp, Bechhofer, and Stevens, 2008). Additional incremental improvements can address this using knowledge engineering practices.

Ontology and Knowledge Learning

The previously noted enormous increase in data, both structured and unstructured, available on the web and in data silos, represents a great opportunity for KG building efforts. But there are known difficulties with some associated data management tasks that have traditionally been done with some degree of handcrafting. However, handcrafting big ontologies and other semantic resources like KGs remains a time-consuming, difficult task. This fact ensures that automated or semi-automated acquisition of ontology from text and structured data remains an active research area with a big payoff potential for populating KGs or building ontologies. Semantic resources like categories of nouns and taxonomies can be learned from texts and even enhanced using automation and machine learning (ML). One way to do this is to extract knowledge from resources on the low end of the semantic spectrum. This idea has been understood for a while (Faure, Nédellec, &, Rouveirol,1998), and efforts have been used dividing the learning process and supporting automation into four different phases: extract concepts, prune, refine, and import or reuse concepts (Mädche, 2005). But the variability of free text can cause a high error rate if automation is based on shallow natural language processing.

While progress has been slow, there have been signs of progress. Inspired by the SW idea of a spectrum of resources. A variety of products can be learned by using a combination of association rules, formal concept analysis, and clustering. Some outputs relevant to KGs focusing on ontologies are shown in Figure 2.

The layer cake model in Figure 2 shows a multistage learning process, starting with terms that are the most basic building block for knowledge learning. Extracted terms can be used to feed into a higher stage using glossaries and thesauri that come with descriptions, definitions and some relations from verb phrases. This process, in turn, provides a basis to define concepts and for generating hierarchies. Along the way, nouns and noun phrases that have the same relations become synonym candidates. All of this helps build a KG information structure that can address questions like “What are the characterizing words, nouns, verbs, and adjectives typically used in this domain?” Ontology learning methods thus provide a way, noisy as it is, to start defining domain knowledge for a KG given a domain

glossary, and it can be more than a simple translation (Bozzato, Ferrari, and Trombetta, 2008).

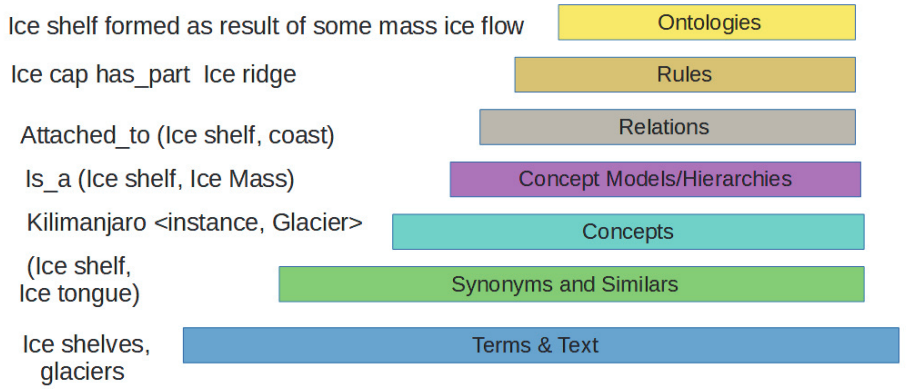


Figure 2: Ontology Learning Layer Cake Hierarchy based on (Buitelaar, Cimiano, and Magnini, 2005

Still, the typical automatically learned semantic resources need humans in the loop. Auto-generated products need to be inspected, validated and modified by domain experts and knowledge engineers before they can be a basis for being accepted, formalized in an ontology, or applied by an application like a KG. Currently, the practical emphasis of such efforts is not to create a final, perfect ontology for a particular domain, but to create reasonable vocabularies first and then incrementally create schema patterns with domain terms and definitions that are good enough (*i.e.*, constrained and defined enough) for people to start using them for publishing data on the Web. A goal is also to support data integration. As with work with different vocabularies and models, a large part of the effort is about harmonizing things to make products useful. And for that we need interdisciplinary teams (Berg-Cross, 2015).

Useful Practices and Some Guiding Principles

Based on varieties of experiences, some improvements in ontological engineering may be suggested to incrementally advance and harmonize semantic resources. The previous examples suggest some ideas, starting with the need to provide unique, persistent and consistent entity identifiers

that are computer processable, but the identifiers should also be relatable and sensible to humans. At a minimum, following FAIR principles, this is needed for findability, but is supplemented when semantic resources are registered or indexed in a searchable resource such as in EnvO. The need for data to be described with rich metadata is also featured in FAIR principles. An incremental approach fits into FAIR's principles for developing and publishing digital material. The use of community standard vocabularies for data should follow FAIR, provide quality, harmonized definitions and associated managed schemas. Vocabularies need to have reliable governance and organizational commitment to FAIR principles and associated ideas like commitment to linked data principles, stable IRI's and associated sensible funding. These are important, as is assessing the readiness of terms, definitions and concepts for use as semantic resources. Incremental improvement efforts need to assess the quality of resources like community accepted glossaries that may exist. Work should leverage and reuse structured data and vocabularies as much as possible. As noted in the work on Cryosphere, progress was enabled by starting with existing vocabularies that were already studied and had some harmonization underway.

Making domain assumptions explicit over time is an important enhancement as we move up the semantic spectrum. References to a foundational ontology, such as done in the OBO Foundry, is an example (Smith et al., 2007). Community efforts like the OBO Foundry represent other basic practices and include principles that are simple, yet important, such as ensuring a stable URI for each refined concept. This goes beyond URL mintings made by individual projects. Efforts like the OBO Foundry involve a community commitment to maintaining a place for managed concepts that assures access in perpetuity.

Broadly, methodologies like eXD or UPON Lite represent useful starting points for projects. The UPON Lite methodology is notable in that it supports the rapid prototyping of trial ontologies that can be extended more easily by enhancing the role of domain experts (using spreadsheets) rather than the constant need for knowledge engineering experts (De Nicola, and Missikoff, 2016). The general advice is to address one modeling issue at a time. For KG development, a useful guideline concerns re-engineering ontological patterns using transformation rules. These can be applied to create a new, target ontology starting from elements of a source model (Blomqvist, Hammar, and Presutti, 2016).

Guidelines also exist for refactoring an ODP. Some concepts may be defined as a hybrid or mosaic of ideas containing a mix of more fundamental concepts. Such marbled concepts have to be decomposed into more rationalized component parts or subtypes since termed concepts should be orthogonal. A guiding principle is that if overlap exists among termed concepts, then there is more than one concept in play. In addition to making use clearer, refactoring affords an opportunity to combine distinct concepts in useful ways. The eXD method provides rules to transform an existing ontological piece, say expressed in OWL DL, due to a requirement change. A typical example relevant to KG development is when a KG is initially populated with individual instances and then advances the organizational use of class structures. Another form of enhancement is moving from object properties to classes (Kasri and Fouzia, 2016).

It is worth noting that ontological methods like eXD require community involvement in the form of interdisciplinary teams. A central emphasis is on the need for domain expertise to address the quality curation that is needed for incremental improvements. This often starts with identifying who are the interested parties needed to improve some collection of semantic resources for some purpose. A typical driver is the need for better data interoperability within a domain. As a bonus along the way, domain experts may provide a seed definition that can expand into new areas of work.

As semantic enhancements are considered, some very general concepts about quality apply to semantic resources across the spectrum, including those such as KGs that make use of many if not all parts of the semantic spectrum. They should not be traded off without thought as we move from one level to another, or as part of efforts to harmonize resources across the spectrum. These concepts are similar to criteria that judge the quality of ontologies and follow from Gruber's (1995) original list including Clarity, Coherence, Extendibility and Minimal encoding. Gómez-Pérez (1996) suggested related criteria that overlap the first three of Gruber's criteria; namely, Consistency and Conciseness (*i.e.*, definitions need clarity and should minimize ambiguity by being expressive in few words). Gómez-Pérez adds that definitions should be complete is some sense and that over time any revisions should capture some core essence of what is known about the real world as part of some finite structure and system. Gómez-Pérez also adds the idea of Definitional Sensitiveness; that is, a definition's core should

be stable in the face of small changes. Some of these qualities are expanded as follows:

a. Clarity: the concepts in a semantic resource should be defined in a formal way that communicates the intended meaning of defined terms as understood by a domain community. Definitions should be brief, with objective necessary and sufficient conditions. If the scope of a defined concept is changing and becoming more formal, that should be documented.

b. Coherence: concept definitions, especially the formal aspects, should stand up to rational analysis. For example logical inferences should make sense and be consistent with the overall domain understanding. It follows the advice – "first do no damage." Wikipedia concept entries have often been the start of a knowledge base. However, based on experience, we know there are consistency questions about these meanings since Wikipedia represents a loosely governed heap of diverse material. While Wikipedia remains a source to start with, it is often better to consider controlled sources that can supplemented by assembling definitions and actually analyzing terms from relevant domain vocabularies which are likely to be stable. This was the case with work on a Cryo vocabulary. A concept does not stand alone, and its definitions and inferences should make sense in light of the definitions and inferences of related concepts. Therefore, methods are needed to reach agreement on conceptualization across concepts. A starting point can be an agreement that something like an ontology or ODP can be assembled by anchoring their concepts to a local, harmonized glossary of terms and a set of agreed upon relationships between them. From this one may consider cross-domain integrating or bridging concepts that meaningfully relate concepts between domains as needed for a wide-spanning KG.

c. Extensibility/Scalability: a suite of semantic resources such as a glossary, KG, ODP or ontology should design in and anticipate expansions and extensions to address likely, new requirements or as semantic resources are added. This is a natural part of a change management process. For semantic enhancements this includes consideration of how to scope a domain, what parts of definitions can be axiomatized and, where possible, use formulations in existing quality ODPs and ontologies for this. Any improved knowledge should are captured in competency questions as laid out in the eXD

method. The previously mentioned SOSA pattern is built on an extendable vocabulary that can be combined with other related ontologies or ODPs, such as SSN, to provide a more rigorous axiomatization where needed. Extensibility is promoted by use of such ODPs as a small conceptual foundation with some general concepts that are useful. Examples include "physical-object", "event", "process" and "situation". These can be used as foundations as a step to a systematic foundry and can be adapted over time to work across a range of certain tasks. A good example of a useful pattern is the formalization of Winston's Taxonomy Of Part-Whole Relations for use in ontologies (Shimizu, Hitzler, and Paul, 2018). A guiding quality related to extensibility that is seen in foundational ODPs is that of "minimal semantic commitment" which states that a semantic resource should require the minimal ontological commitment in order adequately to support any anticipated knowledge sharing activities.

d. Coverage: Coverage asks the scoping or completeness question, "Does the graph have all the required information?" Obviously this is a matter of degree, since even lightly formalized knowledge can't practically provide full coverage of a domain or a suite of domains. As with the work to harmonize Cryo definitions, some scope must be considered. Furthermore, even as extensions and refinements are applied, we know that a KG or an ontology is an approximation and not fully correct. This reflects the reality of a trade-off needed between coverage and correctness. Where and how this trade off occurs is likely to be different in each KG (Paulheim, Heiko. 2017).

On a more detailed, lower level, there are good semantic resource management practices to consider such as:

a. Tracking new concepts added into a KG or ontology, and documenting these incremental changes

b. Versioning concept relationships so that interested parties can explore how they have changed over time

c. The need for Memoranda of Understanding (MOU). As with FAIR, explicit group agreement on meaning and commitment to a core family of defined terms that are central to much work. On a practical level, some between group MOUs on this may be helpful.

d. Nascent technologies exist to help control the quality of semantic enticements, especially to ontologies. These include tools to help with alignment between and among semantic resources such as ROBOT (Jackson *et al.*. 2019) which is a generic command-line tool using a Java library. ROBOT performs common ontology and KG supporting chainable tasks including: computing differences between OWL ontology versions, merging, extracting OWL modules, reasoning, and some support for "explanations". A ROBOT template has been adapted to help align SWEET and EnvO as part of the Cryosphere harmonization previously described. There are also tool suites for broad work on semantic resources. For example, the suite of web-based tools that are part of Ontoanimal (*e.g*., Ontofoxp for reuse of terms, Ontorat to edit existing terms, and Ontobeep for ontology comparisons) support iterative, extensible ontology development (He *et al*., 2018). Boomer (Mungall *et al*., 2016.) is another helpful semantic tool. It uses a combined logical and probabilistic approach to translate mappings into logical axioms for merging ontologies. Tools and practices for pattern-based modular ontology engineering have also been developed (Shimizu, Hammar, and Hitzler, 2020). These reflect a portion of the leveraging and promoting of sociotechnical practices as part of a strategy enabling incremental semantics. Light, rapid methodologies, like the UPON lite ontology approach (De Nicola, and Missikoff.2016) can help make the ontologies more available for KG structuring and consider a range of semantic resources growing from domain terminologies to domain glossaries, taxonomies and simple axiomatized relations.

Conclusions

Knowledge graphs employ a wide range of semantic resources, and bringing them together with rich semantics remains a challenge. But as we have seen, there are now good practices and active research areas using incremental semantic improvements to support this rapidly expanding space of KGs. These range from very focused enhancements to representation of a concept to alignment as part of a community schema or agreed upon domain conceptualization or as part of a larger process to harmonize related but disparate domain vocabularies and assimilate these into extant ontologies. We can expect more systematic development to bring various parts of incremental methods together. These could be evidenced in

improvements targeted to ontological engineering methods for alignment and harmonization. As an example, we can expect to see how to use existing semantic relations found in quality ontologies like EnvO to help improve formal definitions from glossary sources. In turn harmonized glossary definitions provide a rich source of material for assimilation into ontologies and for ODP formation. The result over time will be to help semantically unify richly interconnected heterogeneous data using community adopted and agreed upon methods.

Among the targets for continued research is the use of NLP and ML-based extraction from definitions as seeds for ontology improvement and development of better ML automation. Applying ML to build lightweight ontologies is still exploratory, but promising (Wong, 2009) and there still are challenge of learning non-taxonomic relations needed for ontologies from text. But ML is starting to support tasks such as term extraction, conceptualization and enrichment using reverse engineering, schema mapping, data mining, and ML. Enhanced semantics may also follow from work on ontological subsumption, mapping and ontology matching and handling similarity and analogy models, as well as bottom-up semantics from ML approaches and their symbolic ontology models.

Hopefully, technical and user centered methodological improvements will lead to more community input and involvement. Community involvement is necessary to guide the development and refinement of any semantic resource, including KGs. A core group may offer starting points and facilitate development, but the conceptualization of a domain along with its vocabulary belongs to the community of domain and interdisciplinary experts working with knowledge engineers. This is needed to show how extracted concepts can be grouped, related, and subdivided according to their human-understood semantics in context.

Future extensions to look forward to as part of revised and agreed upon methods include those needed to harmonize and axiomatize entity definitions and extensions using the best ontological engineering practices. But it is likely that we will rely on human agreements for a long time, using devices and helpful artifacts such as community schemas and consistent patterns for resource alignment.

References

Alexopoulos, Panos. *Semantic Modeling for Data*. O'Reilly Media, 2020.

Amith, Muhammad *et al*. "Friend of a Friend with Benefits ontology (FOAF+): extending a social network ontology for public health." *BMC Medical Informatics and Decision Making* 20.10 (2020): 1-14.

Asprino, Luigi *et al*. "Pattern-based Visualization of Knowledge Graphs." *arXiv preprint arXiv:2106.12857* (2021).

Berg-Cross, G. "The GeoVoCamp Workshop Experience and Ontology Design Pattern Development." in Narock, T., and P. Fox. Eds. *The Semantic Web in Earth and Space Science. Current Status and Future Directions* 20 (2015): 171.

Berg-Cross, G and Vardemann, C. Semantic Harmonization: Illustrating Harmonization Levels, Winter ESIP Meeting, Januarym 2020.,

Brickley D, Miller L. FOAF (Friend of a Friend); 2000. http://www.foaf-project.org/. Accessed 28 June 2021.

L. Bozzato, M. Ferrari, and A. Trombetta. Building a domain ontology from glossaries: a general methodology. In A. Gangemi, J. Keizer, V. Presutti, and H. Stoermer, editors, SemanticWeb Applications and Perspectives, SWAP 2008, volume 426 of CEUR Proceedings, 2008.

Buitelaar, Paul, Philipp Cimiano, and Bernardo Magnini. "Ontology learning from text: An overview." *Ontology learning from text: Methods, evaluation and applications* 123 (2005).

Buitelaar, Paul: Overview *"Ontology Learning - Some Advances"* http://ontologforum.org/index.php/ConferenceCall_2017_03_01

Blomqvist, Eva, Karl Hammar, and Valentina Presutti. "Engineering Ontologies with Patterns-The eXtreme Design Methodology." *Ontology Engineering with Ontology Design Patterns* 25 (2016): 23-50.

Choudhari, J. and Suman, U. 2010. Iterative Maintenance Life cycle Using eXtreme Programming. In Proceedings of the International Conference on Advances in Recent Technologies in Communication and Computing (Kottyam, India, October 15 - 16, 2010). ARTCom-2010. IEEE Computer Society, 401 – 403

Dalalah, Ahmad. "Extreme Programming: Strengths and Weaknesses." *Computer Technology and Application* 5.1 (2014).

Darrin, M.A.G. and Devereux, W.S., 2017, April. The Agile Manifesto, design thinking and systems engineering. In *2017 Annual IEEE International Systems Conference (SysCon)* (pp. 1-5). IEEE.

De Nicola, Antonio, and Michele Missikoff. "A lightweight methodology for rapid ontology engineering." *Communications of the ACM* 59.3 (2016): 79-86.

De, Suparna, Yuchao Zhou, and Klaus Moessner. "Ontologies and context modeling for the Web of Things." *Managing the Web of Things* (2017): 3-36.

DiGiuseppe, Nicholas, Line C. Pouchard, and Natalya F. Noy. "SWEET ontology coverage for earth system sciences." *Earth Science Informatics* 7.4 (2014): 249-264.

Duerr, Ruth E. *et al*. "Formalizing the semantics of sea ice." *Earth Science Informatics* 8.1 (2015): 51-62.

El Kassiri, Asmae, and Fatima-Zahra Belouadha. "A FOAF ontology extension to meet online social networks presentation and analysis." *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*. IEEE, 2017.

EnvO May, 19, 20921 release. https://github.com/EnvironmentOntology/envo/releases

European Union, e-Government Core Vocabularies handbook, doi:10.2799/97439, https://ec.europa.eu/isa2/sites/isa/files/e-government_core_vocabularies_handbook_0.pdf (2015)

Faure, David, Claire Nédellec, and Céline Rouveirol. "Acquisition of Semantic Knowledge using Machine learning methods: The System" ASIUM"." *Universite Paris Sud*. 1998.

Michel, Franck. "Bioschemas & Schema. org: a lightweight semantic layer for life sciences websites." *Biodiversity Information Science and Standards* 2 (2018): e25836.

Gangemi. A.Ontology design patterns for semantic web content. In Proceedings of the Fourth International Semantic Web Conference (ISWC-05), 2005.

Gerontas, Alexandros. "Towards an e-Government semantic interoperability assessment framework." *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*. 2020.

Giunchiglia, Fausto, and Ilya Zaihrayeu. "Lightweight ontologies." (2007).

Glacier Hackahon, 2019, https://github.com/Vocamp/Virtual-Hackahon-on-Glacier-topic#readme

Gómez-Pérez, A. Towards a framework to verify knowledge sharing technology. Expert Syst. Appl. 1996, 11, 519–529.

Gruber, T.R. Toward principles for the design of ontologies used for knowledge sharing? Int. J. Hum. Comput. Stud. 1995, 43,907–928.

Guha, R.V., Brickley, D., Macbeth, S.: Big data makes common schemas even more necessary. CACM 59 (2) (2016), http://dx.doi.org/10.1145/2844544

Haller, A., Janowicz, K., Cox, S., Le Phuoc, D., Taylor, K., Lefrançois, M.: 'Semantic Sensor Network Ontology', https://www.w3.org/TR/vocabssn/#intro, accessed January 2018

He, Yongqun *et al*. "The eXtensible ontology development (XOD) principles and tool implementation to support ontology interoperability." *Journal of biomedical semantics* 9.1 (2018): 1-10.

Heflin, Jeff, and James Hendler. *Semantic interoperability on the web*. MARYLAND UNIV COLLEGE PARK DEPT OF COMPUTER SCIENCE, 2000.

Hogan, Aidan. "The semantic web: Two decades on." *Semantic Web* 11.1 (2020): 169-185.

Hull, Richard. "Managing semantic heterogeneity in databases: a theoretical prospective." Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. 1997.

Janowicz, Krzysztof *et al*. "SOSA: A lightweight ontology for sensors, observations, samples, and actuators." *Journal of Web Semantics* 56 (2019): 1-10.

Jackson, Rebecca C. *et al*. "ROBOT: a tool for automating ontology workflows." *BMC bioinformatics* 20.1 (2019): 1-10.

Jupp, Simon, Sean Bechhofer, and Robert Stevens. "SKOS with OWL: Don't be Full-ish!." *OWLED*. Vol. 432. 2008.

Kasri, Soumaya, and Fouzia Benchikha. "Refactoring ontologies using design patterns and relational concepts analysis to integrate views: the case of tourism." *International Journal of Metadata, Semantics and Ontologies* 11.4 (2016): 243-263.

Kim S, Iglesias-Sucasas M, Viollier V. The FAO geopolitical ontology: a reference for country-based information. *J Agric Food Inf.* 2013;**14**(1):50–65.

Kuhn, Werner. "Semantic engineering." *Research trends in geographic information science*. Springer, Berlin, Heidelberg, 2009. 63-76.

Ma, Xiaogang. "Knowledge graph construction and application in geosciences: A review." (2021).

Maciel, Rita Suzana P. *et al*. "Full interoperability: Challenges and opportunities for future information systems." *Sociedade Brasileira de Computação* (2017).

Mungall, Christopher J. *et al*. "k-BOOM: A Bayesian approach to ontology structure inference, with applications in disease ontology construction." *bioRxiv* (2016): 048843.

Niang, Cheikh, Béatrice Bouchou, and Moussa Lo. "Towards tailored domain ontologies." *OM*. 2010.

Noy, Natalya F., and Deborah L. McGuinness. "Ontology development 101: A guide to creating your first ontology." (2001).

Panasiuk, Oleksandra *et al*. "Verification and validation of semantic annotations." *International Andrei Ershov Memorial Conference on Perspectives of System Informatics*. Springer, Cham, 2019.

Presutti, Valentina *et al*. "eXtreme design with content ontology design patterns." *Proc. Workshop on Ontology Patterns*. 2009.

Paulheim, Heiko. "Knowledge graph refinement: A survey of approaches and evaluation methods." *Semantic web* 8.3 (2017): 489-508.

Reynolds , D.(ed.), The organization ontology, World Wide Web Consortium (W3C), 2014. https://www.w3.org/TR/vocab-org/.

Roman, Dumitru *et al*. "The euBusinessGraph ontology: A lightweight ontology for harmonizing basic company information." *Semantic Web* Preprint (2021): 1-28.

Shimizu, Cogan, Pascal Hitzler, and Clare Paul. "Ontology Design Patterns for Winston's Taxonomy Of Part-Whole Relations." *Emerging Topics in Semantic Technologies*. IOS Press, 2018. 119-129.

Shimizu, Cogan, Karl Hammar, and Pascal Hitzler. "Modular graphical ontology engineering evaluated." *European Semantic Web Conference*. Springer, Cham, 2020.

Smith, Barry et al. "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." *Nature biotechnology* 25.11 (2007): 1251-1255.

Taylor, Kerry et al. "The semantic sensor network ontology, revamped." *JT@ ISWC*. 2019.

Wei, J. "Scalability of an Ontology-Based Data Processing System." (2018).

Wilkinson, Mark D. *et al*. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* 3.1 (2016): 1-9.

Wong, W., Y., 2009. "Learning Lightweight Ontologies from Text across Different Domains using the Web as Background Knowledge," PhD thesis, University of Western Australia, School of Computer Science and Software Engineering.