Classification Issues in 2011. Report

Nancy J. Williamson

Faculty of Information, University of Toronto,
140 St. George Street, Toronto M5S 3G6 Ontario Canada
<william@fis.utoronto.ca>

International UDC Seminar, 2011

The third International UDC Seminar was held at the Koninklijke Bibliotheek (Royal Library) in Den Haag, The Netherlands, September 19-20, 2011. The theme of the seminar was "Classification & Ontology." One hundred and forty one delegates from 30 countries were present. Twenty-one papers in 7 categories were presented and 3 poster sessions were held. The papers have been published in the proceedings available from Ergon Verlag and can be purchased online at http://seminar.udc/2011/. Slides and audio recordings can be accessed from the conference programme page.

The setting for the conference was a keynote address entitled "On being the same" delivered by Patrick Hayes (United States). There is a need for semantic insights which better reflect the intertwined ways in which human language use weaves together concepts and descriptions, in a world in which it is commonly asserted that two names or descriptions refer to the same thing, when in fact they are closely related, but not identical. Classification, he says, deals with concepts as opposed to words and raises questions of its ramifications for statements of identity.

The first session of three papers dealt with *The Role of Classification and Ontology on the Web*. All were invited talks and no papers were submitted. Dan Brickley (Netherlands) discussed "Classification, collaboration and the Web of data." He focused on the relationship of subject classification and the Web of data trends around RDF, OWL and SKOS. In particular, he addressed ways in which factual and ontological data might be used along with subject classification and the possibilities this might create for collaboration among maintainers of vocabularies and data sets and in user-facing applications. He indicates that this kind of collaboration is happening but we need to develop some best practice guidance on how the linkages can be created and exploited using practical modern Web tools. His intention was to motivate such collaboration, and he suggested some priorities for the short and medium term. In his talk, Guus Schrieber (Netherlands) was concerned with "Issues In publishing and aligning Web vocabularies." He sees high knowledge value in the application of KOS, such as vocabularies, thesauri and subject headings, to web applications. He calls for methods to publish these systems and clarify their relations and discusses methodological issues in publishing and aligning classification systems on the Web. He explained the basic principles used in building a SKOS version of a vocabulary and illustrated it with examples. In particular, he examined the role of RDF and OWL in this process. Finally, Schrieber presented some examples of how aligned Web vocabularies "can be used to create added value to applications." In the third talk by Thomas Baker (United States), the speaker addressed "The concepts of knowledge organization systems as hubs in the web of data." When KOS are identified, using URIs the KOS function as a hub for accessing resources tagged with its concepts.

The second session of presentations focused on *Classifications and ontologies on their own terms.* In the first paper Barbara Kwaśnik (United States) examined "Approaches to providing context in knowledge representation structures." She selected two cases in which the classification structure faced a number of challenges. The first case describes one American university's dynamic structure and curriculum addressing the question "How can we deeply reflect the complex world of knowledge and practice at a modern university?" As knowledge has grown and developed, its structure has become significantly al-

tered. Lines among disciplines have become blurred and overlapping. Interdisciplinarity abounds, affecting departmental divisions, curriculum content and disciplinary perspectives. Kwaśnik provides a wealth of examples including: Leonardo da Vinci (he is an artist and an engineer); cultural and feminist studies, and areas such as ethics and forensic science. The latter reaches into such areas as anthropology, entomology, chemical analysis, and linguistics. In summing up this analysis, Kwaśnik concludes that "we need better ways of representing and interweaving threads that make up … programmes of study, while at the same time preserving the original understanding of each thread's intellectual home." In the second case she explores "a shifting concept" that she refers to as "living together." This concept has changed over time in its denotation and connotation and raises the question: "How does a classification or ontology designer deal with a subject that is difficult to nail down precisely?" This question is viewed through the lenses of the Dewey Decimal and Library of Congress classifications with analysis of the positioning of the term "cohabitation" in various locations and relocations of the topic. Findings indicated that location was sometimes "apt and useful" while in other cases the classification has "clung to an outdated contextualization." She concludes that universities have used similarly structured curricula for decades, while bibliographic classification schemes evolve and change but still have difficulties. The result is combinations, cross referencing and multiple classes which seem somewhat makeshift at best. Further, she suggests that what is needed is "a way of reflecting the rich dimensionality of subjects as they change over time and reveal different facets in the light of particular context."

The second paper in this session, presented by Richard P. Smiraglia (United States), Charles van den Heuvel (Netherlands) and Thomas Dousa (United States), describes "Interactions between elementary structures in universes of knowledge." It proposes an elementary theory of knowledge based on the structure of knowledge rather than on the content of documents. They explore "three aspects of elementary structures of knowledge that are critical for mediating between the universe of knowledge and the universe concepts, taking *UDC* as their point of departure." They start by considering the components of the elements of structure – the elements and ensembles followed by the order and character of relations as exemplified by classifications as artificial languages and by interaction between elementary structures–*UDC* and other knowledge structures. In the third paper Emad Khazree and Xia Lin (United States) are "Demistifying ontology" The authors take a broad view of the subject and explore attempts to compare the different approaches to organizing them into a model to facilitate collaboration and attempts to clarify the different communities by providing levels of formality, complexity and semantics. Among the systems involved are taxonomies, classification and thesauri.

The third session of three papers focuses on *Classification Meets the Web*. The first paper, prepared by four authors - Daniel Kless, Simon Milton, Edmund Kazmierczak (Australia) and Jutta Lindenthal (Germany) is entitled "Ìnteroperability of knowledge organization systems with and through ontologies." The authors point out that there are differences between the modern ontology and the older types of data modelling such as classification schemes and thesauri. Many efforts had been made to establish interoperability between the types. While the idea is supported, in practice, these efforts are impeded by the absence of standards or guidelines for vocabulary control of ontologies. This paper investigates the interoperability of traditional KOS with ontologies. An introduction to interoperability analyses the function of ontologies using the W3C standard OWL. Some of the issues raised as described in Semantic Web standards are: 1) the different understandings of what an ontology is; 2) the different ontology structures embedded in OWL; and 3) the unaddressed terminological control in ontologies. They set out problems to be faced but also conclude that "ontological principles show great potential for application to existing KOS. Improving their interoperability in terms of easier combination and integration with similarly constructed KOS improved certainty in interoperability, improved search expansion and reduced maintenance." However an initial re-engineering of the KOS would be required. A second paper in this session by Vincenza Maltese and Feroz Farazi (Italy) looked 'Towards the integration of knowledge organization systems with the linked data cloud." The authors cite the importance of sharing and integrating data sets into giant networks of interconnected resources, enabling different applications to interoperate and share their data. In this paper they: 1) highlight potential problems that could arise if purpose and semantics are not taken into account; 2) make clear how difference in purpose is reflected in totally different semantics; and 3) and identify the need for an algorithm to translate from one semantics into another as a preliminary step to integration of ontologies designed for different purposes.

The third and final paper in this section was "Classification and reference vocabulary in linked environment data" prepared by Maria Ruther, Joachim Fock, Thomas Schultz-Krutisch and Thomas Bandholtz (Germany). This paper describes the information system of the German Federal Environment Agency (UBA). The system involves a library together with numerous web systems. The back bone of the system is a classification scheme enhanced by a reference vocabulary which consists of a thesaurus, a gazeteer and a chronicle. Over the years the classification system has been less involved and emphasis has involved the use of the reference vocabulary indexing and full text search. Bibliographic items are no longer classified directly but are assigned to thesaurus terms and the terms are classified. Since 2010 they have been using a linked data system which links bibliographic and observation data with a thesaurus with the classification being visited by inference. The quality and feasibility of an unambiguous classification of thesaurus terms is being questioned. The paper explores the various relationships and their functions. Gradually the classification has faded from the picture. However, in the conclusion the authors comment that "fortunately the classification has survived" and the plan to revive those old classes. As a first step, providing a thesaurus in conjunction with the classification (using Qvoc) implements a crucial prerequisite on the web so each descriptor and each class can be referenced in a resolvable way.

In session 4, two papers looked at *Classification and Ontology of Special Subjects*. Andrew Buxton (United Kingdom) asked the question "Ontologies and classifications of chemicals:can they help each other?" Buxton look first at the enumeration of chemicals in *UDC*. He indicates that the topic ought to be suitable for a synthetic classification but there are problems such as length of numbers, skill needed in construction and the impracticality of construction class numbers for subjects such "steroids" or "neucleric acids." Examples from *UDC* are given. Turning to ontologies he identifies three systems Chemical Entities of Biological Interest (ChEBI), REX, an ontology of physicochemical processes, and FIX that consists of two ontologies, methods and properties also available on the OBO site. Finally he addresses the question "How can classifications help ontologies?" He describes three ways: by providing a readymade hierarchy; by providing a notation; and by providing terms from other disciplines. Conversely he addresses the question "how can ontologies help classification? This he considers under four topics by providing hierarchies; by providing a subject index; by providing access through roles and access via alternative hierarchies. Both provide a structure of concepts but classification does this only while ontologies information about the concepts and their relationships. Each alone has its merits but there are advantages in combining the best features of both. In the second paper in this group Wolfram Sperber (Germany) and Patrick D.F. Ion (United States) discuss "Content analysis and classification in mathematics." They state that in mathematics, classification is not the only approach to content analysis. "Keywords, reviews, summaries, citation analysis and, also in future formula analysis, are important methods of analysing and finding mathematical knowledge." This paper looks at various approaches. It begins with a short review of the history of reviewing journal in mathematics and the Mathematical Classification Scheme (MSC). This is followed by a detailed analysis of MSC including its classes and levels, relations, flexibility and completeness, updating and versions, retrieval and acceptance. Also there is a brief content analysis of two journals ZBMATH and MathSciNet. Finally it looks at the possibilities for future analyses, including a faceted structure for MSC and formula analysis as specific dimensions of content analysis. The authors conclude that the future will require the redesign of MSC and see semantic web analysis as a useful starting point.

Session 5 of the seminar contained three papers on "*Categories and Relations: Key Elements of Ontologies.*" Roberto Poli (Italy) focused on "Ontology as categorical (sic) analysis." There are different kinds of categorical analysis and there are two perspectives– philosophical ontology that focuses on categorical analysis and computer science ontology that aims to create engineering models of reality. Poli is concerned specifically with the former, the analysis of relations that connect categories one to another. The paper begins with a short description of ontology as categorical analysis, followed by the distinguishing of the main groups of ontological categories. Three examples are described: the analysis of a case of paired categories; the basics of the theory of levels of reality; and an in-depth analysis of "temporality"–"i.e. the form of duality linking some paired categories and the relations of superformation and superconstruction that connect levels of reality." In his conclusion he states that "Descriptive ontology is the first layer of any ontological metholology. The categorical features of reality should be extracted from reality itself." In this session, Dagobert Soergel (United States) spoke on the topic "Towards a relation ontology for the semantic

web." In his presentation he noted that semantic web data is structured by computer programs and made available through the Linked Open Data initiative and follows the entity- relationship model encoded in RDF for syntactic interoperability. The semantics of the relationships needs to be made explicit. For this an inventory is needed. In his talk he outlined "a blueprint for such an inventory, including a format for the description/definition of binary and n-ary relations." He drew on "ideas from the classification and thesaurus community … upper level ontological, systems like FrameNet, the Buffalo Relation Ontology and an analysis of linked data sets" (no paper was submitted). In the final paper in this session, Rebecca Green (United States) focused on "Relations in the notational hierarchy of the *Dewey Decimal Classification*. In this paper, the semantic relationship between each of a set of randomly selected classes and its parent class in the notational hierarchy is examined against a set of relational types (specialization, class instance and several flavors of whole-part). "The analysis addresses the prevalence of specific relationship types, their lexical expression, difficulties encountered in relationship types found in *DDC* with those found in other … KOS and compatibility of relationships found in *DDC* with those in a shared formalism like the web ontology language OWL." Details of the results of the analysis are provided in the paper. In final comments Green points out that "a high level of agreement in relationship assignments" was found but that there were several areas where changes would improve the ability to identify hierarchical relationship types in a particular context.

In session 6 of the seminar 4 papers contributed to a discussion of *Modelling concepts and structures in analytico-synthetic classifications.* Ingetraut Dahlberg's paper was entitled "A faceted classification of general concepts." She begins by analysing the term "facet" and arrives at a definition–"A facet (in knowledge organization) is the subdivision of an *ur*-category used to form mutually exclusive classes of relevant concepts." This is followed by a brief overview of the history of the contributions to the development of facet analysis–the work of the Classification Research Group, Ranganathan, Paul Otlet on *UDC* and the contributions of *BC2*. The details of Ranganathan's seminal mneumonics and the amenities of zero level are discussed. "Representing the structural analysis of a freely faceted classification" was presented by Claudio Gnoli (Italy), Tom Pullman (United Kingdom), Phillippe Cousson (France), Gabriele Meril (Italy) and Rick Szostak (Canada). Based on the theory of Freely Faceted Classification (FFC) drafted by the Classification Research Group (CRG) this paper describes the research in an experimental system called the Integrative Levels Classification (ILC). It currently contains approximately 7050 classes recorded in a MySQL database and a first fixed edition (ILC1) has been released on the Web. Here the structural elements of the notational and verbal planes are described. Subsequent sections discuss the display of elements in a Web interface and reveal problems and preliminary solutions for their representation in KOS. The results show the complexity and layering of faceted classifications and identifies additional features not found in other knowledge organization systems that will need special treatment if their potential is to be fully exploited. Vanda Broughton described the use of "Facet analysis as a tool for modelling subject domains and terminologies." She states that the features of faceted classification "provide an excellent basis for the general conceptual modelling of domains, and for the generation of KOS other than systematic classification." This is demonstrated by a faceted approach to many web search and visualization tools, and the emergence of a facet based methodology for the construction of thesauri. Current work on the *Bliss Bibliographic Classification* is investigating ways in which the full complexity of faceted structures can be further utilized. The paper sets out the organizing principles of faceted structures with examples from *BC2*, describes *BC2* thesaurus software and the web enabling of *BC2*. It looks briefly at the use of faceted classification in exchange formats such as XML. The work so far has identified six conclusions and the author believes that a number of research question related to the semantic Web could be "tackled through the medium of facet analysis."

The final paper in this session was "Analytico-synthetic approach for handling knowledge diversity in media content analysis" by Devika P. Madalli and A.R.D. Prasad (India). Using a faceted indexing method, illustrations are used to demonstrate facet analysis and synthesis for use in annotations for media content analysis in a "Living Knowledge" project funded by the European Union (EU). The project is intended to study the effects of diversity and time on opinions and bias. Included is a brief discussion of facet analysis and faceted classification in general. This is followed by a description of the application of faceting to Media Content Analysis (MCA) services; a service that analyses media reports in different formats. The analytico-synthetic approach is applied to a multidimensional faceted representation from the domain of political science.

In session 7, the final session of the conference, there were 2 papers on *Transforming and Extending Classification Systems.* Joan S. Mitchell, Marcia Lei Zeng (United States) and Maja Zumer (Slovenia) gave a presentation on "Extending models for controlled vocabularies to classification systems: Modelling *DDC* with FRSAD. The authors are considering whether or not the FRSAD (Functional Requirements for Subject Authority Data) conceptual model can be extended beyond controlled vocabularies to model classification data. The purpose of the research is "to test the applicability of the FRSAD model for classification data, and as a springboard for a general discussion of issues related to the use of FRSAD for the representation of classification data." The paper describes the FRSAD model and the *DDC* case study including examples. The authors conclude that the study meets their expectations and identifies two proposals for future work : the extension of *DDC* modelling to investigate issues related to translations of *DDC* and the mapping of controlled vocabularies; and experimentation with modelling other classification schemes, e.g., LCC and *UDC*. The final paper of the conference, "Transformation of a legacy *UDC*-based classification system: Exploiting and modelling semantic relationships" by Fran Alexander and Andy Heather (United Kingdom) reviewed a project to remodel and unify diverse BBC archive schemes including a large *UDC*-based classification, Lonclass, as a part of the BBCs Digital Media Initiative (DMI). The aim is to replace the existing structures with a faceted structure of classification. The introduction gives a brief history of the BBC Archives classification schemes, states the DMI objectives and defines terms. The use of facet classes as a basis for ontological relationship modelling is discussed along with the use of a network-based model as a pre-cursor to moving to a rule-based ontology and the complex methodology used to decompose and reassemble the classes is detailed with examples. Although challenging, this process identifies ways of preserving the high precision semantics of bibliographic classifications for use as a foundation for natural language-based retrieval and integration into ontologically expressive formats such as Resource Description Framework (RDF). Further work is needed on methodology and standards similar to those for SKOS and RDF.

There were three poster sessions held during the seminar and short papers on these topics are included at the end of the proceedings. Included are presentations on: "The evolution of knowledge and its representation in classification systems" by Andrea Scharnhorst, Amila Akdag Salah, Krzyszrtof Suchecki, Cheng Gao (Netherlands) and Richard Smiraglia (United States); "Visualizing universes of knowledge: Designs and visual analysis of the *UDC*" by Charles van den Heuvel, Almila Akdag Salaha (Netherlands); and, "*UDC* as a knowledge framework for building a civil engineering ontology: a practical approach to knowledge representation and visualization," by Ricardo Eito-Brun; Alfredo Calosci (Spain).