Keynote Address: UDC Consortium Seminar, Lisbon, Oct 29, 2015. http://seminar.udcc.org/2015/
This is a lightly revised version of text published in the proceedings: *Classification & authority control: expanding resource discovery: proceedings of the International UDC Seminar 2015, 29-30 October 2015, Lisbon, Portugal*. Edited by Aida Slavic & Maria Inês Cordeiro. Würzburg: ERGON-Verlag, 2015. 248 pages. ISBN: 978-3-95650-124-1. Accompanying slides are available at http://people.ischool.berkeley.edu/~buckland/lisbonudcc2015.pdf

Classification, Links, and Contexts: Making Sense and Using Logic.

Michael K. Buckland, University of California, Berkeley, USA

INTRODUCTION

The title of this Seminar is "Classification & Authority Control: Expanding Resource Discovery" and, within that, the announcement states that:

> Linked data practices and techniques have opened new possibilities in exploiting controlled vocabularies and improving resource discovery. Authority data held in library systems often includes classification schemes. These knowledge structures now have the potential for being shared across the linked data environment. The objective of this conference is to explore such potential, expanding the value and use of classification as an authority controlled vocabulary, from a local perspective to the global environment.

Differently stated: Knowledge organization systems have vocabularies with internal links between terms and now we are challenged to improve resource discovery and selection through the use of external links between terms in different vocabularies.

The theme is well-chosen because it addresses two issues: First, promising technical developments in professional practice; and, second, basic questions concerning what we are doing when we make such links and what the limits of using such links might be. Other speakers and seminar participants are very expert in these technical developments. I have been asked to consider these issues within in a wider context and so I will focus on the second issue, on basic questions.

We are here to improve resource discovery in the spirit of Paul Otlet's grand vision of universal access for all media, for all topics, and for all people everywhere. This dream lives on in the Semantic Web and in this Seminar's agenda. It is our vision and we should remember why: our purpose is to empower people to learn, to know more, and to understand better than before. And, following Otlet's admonitions, I will start with standards.

BEGINNING WITH FRBR & FRSAD

Functional Requirements for Bibliographical Records (*FRBR*, 1997) is based on three groups:

Group 1 entities are defined as the products of intellectual or artistic endeavors that are named or described in bibliographic records and are of four types: *work*, *expression*, *manifestation*, and *item*. The last two are material, tangible objects, for example, respectively, a publisher's edition as a set and or an individual copy of a book in that edition. The first two, *work* and *expression* (a version of a *work*) are conjectured, metaphysical entities that are not perceptible to our senses. *Works* and *expressions* of them can be operationalized only by descriptions of what we imagine them to be or, more pragmatically, inferred inductively from sets of *manifestions* and *items*. I like to use the term *document* as an informal shorthand for Group 1.

Group 2 entities are those responsible for the intellectual or artistic content, the physical production and dissemination, or the custodianship of the Group 1 entities: person, corporate body, and family. In short, whoever is credited as *creator*, although we know that the realities of creation are more diffuse than cataloging theory concedes (McKenzie, 1986).

Group 3 entities represent an additional set of entities: the subjects or topics of *works*. These are addressed in a separate standard FRSAD.

Functional Requirements for Subject Authority Data (*FRSAD*), 2010.

FRSAD uses two types of entity: *Themas* and *Nomens*.

Thema is a synonym for topic, what a *work* is about or of or on. The examples given of what could be a *thema* include physical objects, conceptual entities, and events. A *thema* could be anything sensed, perceived, or imagined. A good technical term for anything sensed, perceived, or imagined is *phenomenon*, one can say that *themas* are phenomena.

Nomen is a synonym for name and so here means a name of a phenomenon identified as a topic (*thema*). The name can be in any form: subject heading, classification numbers, ontology node, category code, keyword, tag, etc. *Nomens* are names ("nominations"), assigning *nomens* is a language acts. Languages can be thought of as composed of names related grammatically.

A *vocabulary* is a set of names used within a language. When used for description in resource discovery a vocabulary is commonly "controlled" into preferred and non-preferred terms to accommodate semantic equivalence (synonyms), inclusion (hierarchy), other relationships (see also).

Relating (linking) a *nomen* in one language (vocabulary) to a *nomen* is in another, different language is called "mapping".

Language, especially vocabulary, evolves within social groups (domains, fields of discourse) and evolves continuously. It follows that mapping between different domains is a mapping from a name in one context to a name *in a different context*. This is important in several ways: Humans reason, more or less. We make sense within whatever context we are in. We make decisions. Reason is the ability to make sense, to take decisions, and to set up stable theories. Reason is importantly different from rationality which is the ability to follow rules. Humans are not strictly rational (hyper-rational) in the way that logic, algorithms, digital computers, and set theory are. This difference between reasoning (making sense) and strict rationality (following rules of logic) is important for our Seminar theme.

So our Seminar theme can be viewed at two levels:
Application: How best to combine links and vocabularies for resource description and discovery, and
Insight: What can be said about relationships between phenomena, names, links, and human understanding?

In other words: how can we provide better service; and how can we understand better what we are doing.

SUMMARY OF ASSUMPTIONS

In addition to what has just been said about terminology and language:

1. Learning, knowing, and understanding constitute a part of how we live, so Documentation (by whatever name) is unavoidably a form of cultural engagement. Why support resource discovery if not to facilitate learning more and/or a changed understanding? (Buckland 2012b).
2. Our systems are full of links of many kinds, including subject indexes, syndetic structures in thesauri, search term recommender services, as well as "linked data" in the sense of Linked Open Data.
3. There is a significant difference between making sense and using logic and a tension between language and logic.
4. Probabilistic methods can be useful in this complex and unstable world.
5. The challenge underlying this Seminar is to embrace the full expressive power of language, to acknowledge the cultural complexity of our environment, and yet use formal, rational tools that are powerful.

HOMMAGE TO PAUL OTLET (1868-1944) – AND LUDWIK FLECK

At any UDC Consortium event it is proper to pay homage to Paul Otlet and his concern with building a universal resource discovery system based on standards, compatibility, systems, and

machinery (Buckland, 2007). His very first paper on this topic in 1892 is a good start. Entitled "Un peu de bibliographie" (Something about bibliography), it is a plea for collective action in analyzing, cataloging, and classifying facts for "the creation of a kind of artificial brain by means of cards containing actual information or simply notes of references" and, relevant to this Seminar, he includes a plea for vocabulary control through:

> ". . . a careful arrangement of its nomenclature . . . because of its conciseness and the influence it would have on the creation of a scientific language *ne varietur* of which the social sciences are particularly in need. This nomenclature, which usage would soon shape and stabilise, could also be used for the classification cards in the catalogue. It would thus permit the creation of very practical links . . ." (Otlet, 1892/1990:19).

However, Otlet was introduced very well at the last Seminar by Boyd Rayward (2013). My impression is that Otlet regarded links and vocabulary control as a solved problem: You use the Universal Decimal Classification for all arrangements of documents and you provide natural language indexes to it.

I think that we might understand Otlet better if we paid more attention to his contemporaries and their ideas (Buckland, 2012a). (Otlet was an enthusiastic collector of ideas. I think he would approve.) For this year's Seminar theme, I draw attention to Ludwik Fleck (1896-1961). Fleck was a Polish microbiologist. He was thirty years younger than Otlet, but his key work, *Genesis and development of a scientific fact*, was first published in 1935, the same year that Otlet published a summary of his theories, *Monde*, and one year after Otlet's encyclopedic, *Traité de documentation* (Fleck, 1935/1979; Otlet, 1934, 1935; Sady, 2012).

Fleck argued that a text has to be understood in relation to three entities: the writer, the text, and the cultural habits and cultural context of the writer (*Denkstil*, *Denkkollectiv*). And when a text is read it is necessarily read with the cultural habits and cultural context of the reader. So there is a double Fleck effect: Not only the writer, the text, and the *writer's* cultural context, but also the reader, the text, and the *reader's* cultural context. Difficulties arise from differences between the two cultural contexts. We understand ancient, medieval, and renaissance authors with difficulty because the writers' knowledge and their ways of thinking are so different from ours. And those writers would have difficulties understand our current writings.

For similar reasons, Fleck was critical of factual encyclopedias which were central to Otlet's vision, shared with Wilhelm Ostwald and H. G. Wells, for a universal encyclopedia with many small factual entries ("monographs") linked together to constitute a "World Brain." Fleck's view was that brief factual entries in a reference work were necessarily unsatisfactory because the reduction to brief facts removed explanatory context necessary to make adequate sense. Hence, Fleck's insistence that cultural context is important for making sense is subversive of Otlet's modernist, global vision. I see the tension between Fleck and Otlet, between the local and the global, as fundamental to our Seminar theme.

In a narrow interpretation, linked data uses standards and practices (URI, RDF, OWL) that allow machine operations within the Semantic Web vision of a World Wide Web of data in which machines could discover and process resources independently of humans. A broader view of linked data would include all forms of links likely to be useful in documentation systems, including authority control and mapping between different vocabularies.

VOCABULARIES

Sometimes topic headings are hard to predict but easy to understand once found, e.g. Hand-to-hand fighting, oriental, in motion pictures. (A former Library of Congress Subject Heading for Kung-fu films). Sometimes they are hard to understand as well as hard to predict, e.g. HS 847120: Digital auto data proc mach contng in the same housing a CPU and input & output device. (*International Harmonized Commodity Classification* heading for *Computer*).

> And there are many names in different vocabularies for the same *thema*, e.g. for automobiles:
> 629.331 (Universal Decimal Classification)

PASS MOT VEH, SPARK IGN ENG (US Federal Import/Export statistics)
TL 205 (Library of Congress Classification)
180/280 (US Patent classification)
3711 (Standard Industrial Classification)

etc., etc. Nobody can be expected to know all the names in all the vocabularies used in potentially useful resources. Hence the need for links and for recommender services when links are absent or inadequate.

MAPPING: LINKS BETWEEN VOCABULARIES

A link, by definition, leads from one point to another point elsewhere. So, necessarily, a link between vocabularies leads from one context to another. Otlet, the enthusiast, would see links as making a connection, a positive achievement. Fleck, I suspect, would remind us that moving from one context to another is problematic for meaning and sense.

We can start with a simple example from the subject index to the first edition of the Dewey's decimal classification in 1876: Railroads 385 where the reference implies equivalence. "The number after it is its class number and refers to the place where the topic will be found," explained Dewey (1899: [405]). It is a link between terms in two different languages. Such a mapping is necessary whenever an artificial language is used. Starting with a classification, then adding natural languages avoids the usual problems of vocabulary control. We could call Dewey's "relative" index a search term recommender. We can mention some elaborations:

(1) Dependence on context. The sixth edition of Dewey's *Decimal Classification and Relativ Index for Libraries, Clippings, Notes, etc*. of 1899, used Railroad to illustrate that the most suitable link varies according to the context ("in different connections" (Dewey, 1899: 10)):

Railroad	architecture	725
	corporations	385
	engineering	625
	law	385
	travel	614.863

(2) Direct mapping between more than two vocabularies.

The use of standardized indexing languages can be recommended but not mandated and might not always be entirely desirable. So there are always multiple different vocabularies relevant to any given topic. Most of these vocabularies will be unfamiliar for any given user, so helping searchers to use unfamiliar vocabularies by providing links in the form Railroads 385 becomes increasingly needed. The Unified Medical Language System (www.nlm.nih.gov/research/umls/,UMLS) is an outstanding example of direct mapping between many health and biomedical vocabularies and standards to support interoperability. Creation of such mappings by humans is difficult and expensive and increasingly so as the number of different vocabularies increases. And the results are necessarily obsolescent because knowledge and word usage (both part of Fleck's cultural context) evolve as time passes (Buckland, 2007, 2011).

(3) Indirect mapping between more than two vocabularies.

The number of links required can be greatly reduced if terms in second and subsequent vocabularies are mapped to the first to provide a star network. A disadvantage is that most links are indirect: to the first vocabulary and then a second link to the target vocabulary. Since links are rarely perfect semantic matches, the additional links are likely reduce the exactness of match. Something is lost in translation.

The multiple subject indexes in different natural languages to the Universal Decimal Classification all link to the same set of class numbers, so, although it was not the intention, terms in one natural language can lead (via the class number) to corresponding terms in other natural languages using the classification number as a "pivot" or switching language (also known as

intermediate lexicon) to map presumed synonyms across multiple languages to each other through a star network. See www.udcc.org/udcsummary/php/index.php.

331.2 Salaries. Wages. Remuneration. Pay	(English)
331.2 Salajroj. Rekompenco. Enspezo. Lukro	(Esperantoi)
331.2 Salários. Ordenados. Remuneração. Pagamento	(Portuguese)
331.2 Gehälter. Löhne. Lohnzulagen. Honorare	(Deutsch)

(4) Unfamiliar documentary languages. The link Railroads 385 leads from a familiar term to an unfamiliar term. Experienced searchers know that familiarity with a classification or indexing vocabulary improves both effectiveness and efficiency in searching. People search less effectively and less efficiently when using documentary languages that are not familiar to them. The expanding range of the Internet is like getting access to a much larger reference library. However, although a greatly increased range of resources has become available, these additional resources are likely to have different, less familiar documentary languages. The number of accessible resources has increased, but, also, the proportion of available resources with unfamiliar metadata has increased and it has become more difficult to search effectively or efficiently. (Consider our *automobile* example above. How many people know all those search terms?).

Probabilistic mapping

Fortunately, probabilistic methods can help, as we found in studies at Berkeley based on the important work by Ray Larson on "classification clustering." Larson showed that a sample of catalog records could be used as a training set to estimate how individual Library of Congress Classification (LCC) numbers were statistically associated with words found in titles and in Library of Congress Subject Headings. From that he was able to predict with some reliability from the words in a book's title what LCC number would be assigned to it (Larson, 1991, 1992).

A system that can predict assigned classification numbers can also be used as an economical automatic classifier or, better, as a human classifier's aid. Equally the same capability could be used as a search term recommender for anyone wanting to know suitable LCC numbers for any given topic. Further, this methodology is good for the mapping of *any* two vocabularies (whether natural language words or classification numbers) if a suitable training set can be found. This versatile and economical method for creating "see" references like Railroads 385 from familiar terms to unfamiliar ones can be frequently regenerated using more up-to-date training sets to reduce the effects of obsolescence. A series of studies exploring these issues was undertaking at Berkeley. (Buckland, et al. (1999) provides a concise summary. Petras (2006) provides a detailed explanation. This approach is now also called "instance-based matching" (e.g., OAEI, 2014)).

COLLECTION VOCABULARY AND DOMAIN DISCOURSE

Any large collection (database, catalog, bibliography) usually covers a range of subject matter and includes specialized subdomains, each with its specialized word usage. Anyone interested in a subdomain needs the specialized concepts and vocabulary of that subdomain, so there should be advantages in having a suitably specialized index that would improve resource discovery in that subdomain within the larger collection. Specialized indexes can be generated by using training sets drawn only from records within that subdomain (Buckland, et al., 2000).

In our first experiment with subdomain vocabulary mapping, we used the INSPEC database and in addition to a general search term recommender based on an unbiased, collection-wide sample, we also picked a separate specialized sample for each of three topical subdomains: Information Science, Biotechnology, and Water Resources to create three different search term recommenders. The first query posed was "Galileo" and the differently-derived search term recommenders yielded different search term recommendations:

The collection-wide index recommended: "*Jupiter*" then "*Planetary sciences*"

The Information Science index: "*Reservation computer systems*" then "*Travel industry*"

The Biotechnology index: "*History*"

The Water Resources index: "*Planetary atmospheres*"

On inspection we found that the first arose from the space probe named Galileo then seeking evidence of water on the planet Jupiter and its moons. The second derived from the Galileo online ticketing system then used by the travel industry. The third recognized an historical name, Galileo Galilei. The fourth was also derived from the Galileo space probe. The recommendations differed greatly, but *each recommendation was valid within its context*!

"Cardiac arrest" in different specialties

Imagine three doctors -- an anaesthesiologist, a drug therapy specialist, and a geriatrician – who each wanted recent literature on cardiac arrest (i.e. heart attack). Cardiac arrest itself is not a MeSH (Medical Subject Headings) heading, so what would be the most suitable MeSH headings to use? The three doctors are specialists. They do not have the same kind of interest in cardiac arrests. As specialists, they inhabit different medical micro-cultures. Each would not be interested in (and might not understand) the specialized literature of interest to the others. Suitably biased training sets can generate specialized search term recommendations for each. See Table 1.

Table 1: *Medical specialty search term recommender results* (Petras, 2006)

Recommended MeSH Headings for query "Cardiac arrest" for three specialties		
Anaesthesiology	Drug therapy	Geriatrics
1. Heart arrest	1. Heart	1. Coronary disease
2. Heart surgery	2. Purkinje fibers	2. Heart diseases
3. Cardiac output	3. Myocardium	3. Crime
4. Respiratory insufficiency	4. Anti-arrhythmia agents	4. Heart
5. Heart attack, induced	5. Arrhythmia	5. Cardiovascular agents
6. Heart	6. Heart conduction system	6. Mitral valve insufficiency
7. Heart diseases	7. Cardiac output	7. Cardiomyopathy, hypertrophic
8. Resuscitation	8. Myocardial contraction	8. Aortic valve insufficiency
9. Coronary artery bypass	9. Anilides	9. Up-regulation (Physiology)
10. Hyperkalemia	10. Heart arrest	10. Pindolol

Each column can be interpreted as follows: In the context of anaesthesiology, recommended search terms for documents relating to "cardiac arrest" are 1. Heart arrest, 2. Heart surgery, etc. Each ranked list appears reasonable for its specialty, but they are remarkably different from each other and they lead, quite properly, to different documents.

More generally, context-specific training sets significantly improve retrieval performance within a subdomain compared with an "unbiased" (collection-wide) training set. They also severely decrease performance if the wrong specialty recommender is used. The more the training set is specific to the context (the specialty) the better the improvement in resource discovery within that specialty and the worse the results if the wrong specialty index if used. There are several ways to define context-based training sets. Using a classification, when available, allows a specialty to be defined efficiently and effectively by choosing classes of interest (Petras 2006).

Personalized contexts

Commercial advertising has become more targeted within data-rich environments with the ideal being the "market of one." So perhaps for resource discovery the ideal would be entirely individualized, personalized indexes: *In the context of your own personal knowledge,* for A see B. An expert reference librarian or teacher would try to do this. It is a logical consequence of shifting

attention from Otlet to Fleck and it is completely different from the traditional practice of one single, collection-based index for all users, which was the only affordable method before digital technology.

The ideal of unique personalized indexes might be difficult to implement, in part because what one knows reflects our complex personal cultural context It is more than and different from a single topical specialty. Nevertheless, the idea is worth considering, especially given the general failure in library and information science to act on the basic principle that what a person already knows is highly relevant to what resource discovery would be most helpful (Konrad, 2007: 517-593, esp. 526). The targeted advertising we receive online based on our past searching and past online purchases provide a precedent. Ron Day has pointed out that "The data—it is me!" (Day 2013; 2014). The situation reminds me of the old saying: Man makes clothes, the clothes make the man.

LINK RELATIONS

Links are seen as logical relations that could be operationalized algorithmically in a Semantic Web sense, but many links express relationships that are vague or unspecified or difficult to use algorithmically. I suggest that all links and mappings are descriptive and could, therefore, be regarded as language acts in a broad sense and that only a subset are recognizable as logical relations. In this view, logical links that can be used algorithmically are a small subset of language links.

Much attention is given to semantic links based on semantic similarity, inclusion/exclusion, and inheritance: vocabulary control (syndetic structure within a documentary language); authority control (links to preferred forms); and mappings between terms in different vocabularies (categories, classifications, ontologies, subject headings, thesauri). However, a broad view of links would include all forms of reference and citation, including the link between a query and a retrieved set. Selection systems work algorithmically, so the internal logic of retrieval systems can also be considered a kind of link: a given query always responds with (links to) the same retrieved set (also known as system relevance) in any stable situation.

Links are commonly in the form of A <sameAs> B. We know that, strictly speaking, this is always false because no two different entities can ever be the same, as Patrick Hayes explained in the 2011 Seminar (Hayes, 2011). In reality A <sameAs> B means that B is an acceptable alternative to A, similar enough for some purpose, and ultimately this relationship is situational. (In English, when quite different alternative options are equally acceptable, one might say, "It is all the same to me." But sometimes these options might not be equally acceptable.)

Thesauri and ontologies use some additional relationships of inclusion, ancestry and hierarchy, but in practice the repertoire is small and limited to relationships suitable for logical inference.

Synonyms in different natural languages

Consider *Alcoholism* (English), *Alkoholismus* (German), and *Alcoholismo* (Spanish). These three words would ordinarily be considered synonyms, but when used as search terms the situation is less simple. They will, of course, find English, German and Spanish language texts respectively, but the discourses found have different concerns and emphases in addition a focus, explicitly or implicitly, on different geographical areas based on their language area or country of publication.

An emphasis on different aspects of alcoholism may be associated with differences in cultural context. These emphases can be inferred from the additional subject headings assigned, which indicate interest in drug abuse and ethyl alcohol in English language discourse, in employment issues and women alcoholics in German language discourse, and in youth and drunkenness as a criminal offense in Spanish language discourse. Spanish and, especially, German books seem to emphasize social and psychological aspects of alcoholism and the English-language books give more attention to medical and psychiatric issues. To the extent that *Alcoholism*, *Alkoholismus*, and *Alcoholismo* are associated with different discourses in their different cultural contexts, they are not equivalent *as search terms*.

FUNCTIONAL RELATIONS

Some links reveal functional rather than semantic relationships. Consider *Pig manure*, *Water hyacinths*, and *Biogas*, which differ greatly in their nature, appearance, and smell. In operational systems these three are likely to be only distantly connected, if at all. They are semantically distant. The *Library of Congress Subject Headings* has two headings, *Biogas* and *Biogas Industry*, with almost no related headings. However, pig manure and water hyacinths are major ingredients for making biogas, so in resource discovery, anybody interested in one is likely to want to be made aware of the other two. Such functional relationships (material relations) are easily identified by examining frequently co-occurring subject headings within a retrieved set.

WHERE: PLACE AND SPACE

"Where" is interesting because there is a dual naming structure that is very be useful for resource discovery: *Place* (a cultural concept) and *Space* (a physical structure). Place names can refer to specific geographical divisions with known boundaries or to areas of interest whose extent may be only vaguely defined. A place may have multiple, unstable names; the nature of the place ("geographical feature type") may change (e.g. from village to town); and the boundaries of the area denoted will be more or less unstable with political changes.

Space can be expressed in terms of latitude and longitude. Spatial description (geo-referencing) can be used to disambiguate different place names and to show relationships between them, and geographical context. Place name lists (gazetteers) linking place names to spatial descriptions (latitude and longitude) like a kind of bilingual dictionary are very useful (Hill (2006), Buckland et al. (2007).

WHEN: EVENTS AND TIME

There is a similar duality of events and calendar time. Not only do events take place in time, but people commonly mention events as a way to identify a point in time. chronological markers]

WHO

Biographical records provide interesting examples. There are unlimited ways in which humans can relate to each other, but in practice databases about persons are largely limited to a small set of familial relations: parent / child; spouse; and sibling. Some prosopographies have an expanded range of relationships. For example, the *Standards for Networking Ancient Prosopographies: Data and Relations in Greco-Roman Names* (SNAP:DRGN) (Standards, 2015) project has fifty-six classes of interpersonal relationships, including <FreedmanOf> and <ExtendedHouseholdOf>. Similarly *AgRelOn*, an Agent Relationship Ontology, has an enlarged range of familial, quasi-familial, and a few other relationships (e.g., murderer, muse, physician) (Deutsche National Bibliothek, 2011). Nevertheless, these expanded repertories are only a small selection of all possible relations. One could not write a biography with such a limited set of relationships. Maybe prosopographies are about persons not lives or they were limited by design to relationships that lend themselves to logical inference.

We can illustrate this limitation by considering a biographical record in the style of the Who's Who genre for another pioneer from the period of Otlet and Fleck, Emanuel Goldberg, who developed an early search engine.

> Emanuel Goldberg, b. Moscow, 1881; son of Grigorii Goldberg; Univ. of Moscow, 1900-04; Ph.D w. Robert Luther, Leipzig Univ., 1906; m. Sonja Posniak, 1907; Assistant, Adolf Miethe, TU Charlottenburg, 1906-07; Prof, Akad. f. graphische Künste, Leipzig, 1907-17; ICA, Zeiss Ikon, Dresden, 1917-1933; Kinamo cine camera, 1921; microdots, 1925; search engine, 1927; Contax 35 mm camera 1932; kidnapped by Nazi SA; refugee in Paris, 1933-37; Laboratory, Palestine, Israel, 1937; d. 1970.

What sense a reader might make of such a record depends on the reader's prior knowledge. The more you know about doctoral study in Leipzig in 1906 or "Nazi SA" the more fully one would

understand Goldberg's experience. Which explanatory resources would be most helpful will be situational, depending the knowledge, language skills, and cultural context of the reader.

In this example there are many proper names that could easily be linked to authority files for resource discovery. But even this brief example includes non-familial relations: <studied at> <studied with>, <assistant to>, <kidnapped by>, <taught at>, <invented>, and <became refugee in>. If we are to be serious about using links for expanding resource discovery then starting with the richness of the human experience is more attractive than a limited set of logical relationships.

In the case of biographical narratives we have experimented with decomposing a life into a series of events, each of which appears to form 4-tuples of values for persons, action/status, place, and time (Buckland & Ramos, 2010). Whether or not that design would prove adequate and robust is less important than recognizing that if we are to understand a person's life it is necessary to understand the contexts (cultural as well as spatio-temporal) and also the full range of what individuals do during their lives. In reality, the list of interpersonal relationships is very long: friend, fellow-student, critic, plagiarist, admirer, co-worker, committee colleague, military comrade, and so on. And people have many relationships to objects (designer, maker, owner, collector, repairer,…) and performances (witness, instigator, victim, chronicler,…).

Using relations

While <sameAs> and its variants are clearly useful for purposes of identity and disambiguation, our seminar agenda of expanding resource discovery should include using descriptive links for the full range of relationships. Maybe the only way to do that is to use natural language or a very versatile universal artificial one, such as the UDC. (See American Library Association, 1997).

We could use natural language A <English: co-worker> B. We could use a classification, maybe A <UDC: 622.33> B, for A was a coal miner with B. Then algorithmic processing could attempt operations on such statements. This implies use of complex inference using probabilistic methods, natural language processing, and/or network analysis after an initial resource discovery to select statements. I suppose ontologies could in theory have unlimited relations. But how well would that work? Perhaps the syntax and relationships of synthetic classification systems could replace the limited logical relations currently in use. If all relations that a computer might encounter and not recognize were defaulted to a generic <is related to>, little has been lost and something may have been gained.

FACETS, LINKS, AND CONTEXTS

Facets, the most fundamental divisions, are central in classification and knowledge organization practice. Because linked data normally have a <sameAs> or similar relationship mappings are ordinarily within the same facet. Likewise, in a library reference collection, the reference works are arranged by facet-specific genres: biography (Who) is separate from geography (Where) which is separate from history (When), and so on. But when we look beyond the main heading in a knowledge organization system (or at the body of an entry in a reference work) we find no such limit to a single facet. Instead we find any or all other facets.

-- A library subject headings will have a main heading, but then commonly a geographic subdivision, a chronological subdivision, or a form designation, and, sometimes, a personal name.

-- A place name gazetteer will have a place name as a main heading, but then a geographical feature type, and spatial markers for latitude and longitude. It could have a note of when that name was in use and/or mention of an especially significant associated person or historical event.

-- A time period directory would have a period name as heading, qualified by what kind of period or event it was, time markers (calendar time), and where that period or event occurred.

-- A biographical dictionary will be arranged by personal name, but, as in the Goldberg example above, followed by multiple instances of activity, date, other persons, and locations (e.g. "b. Moscow, 1881; son of Grigorii Goldberg", etc.).

Instances will vary greatly. The important point is that any main heading in a bibliography or catalog and any entry in any facet-limited reference work is likely to have qualifiers or explanations using other facets. So same-facet links can be made across different facet genres. Diagram 1 shows what one might expect, with lines connecting instances from different facets: When, Where, Who, and What. We see the same effect in pre-coordinate systems such as the Library of Congress Subject Headings (LCSH) and synthetic systems such as the Universal Decimal Classification (UDC).

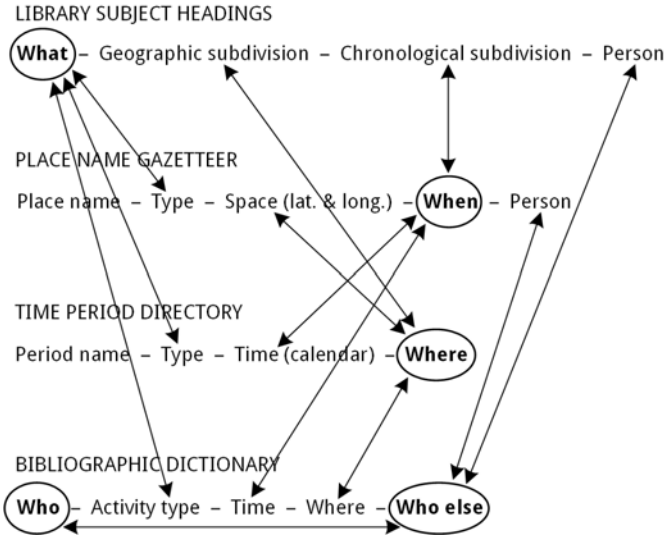

Figure 1: Same-facet links across different facet genres.

There may be reasons for the sequence of the facets in each row, but if, for the purposes of discussion, we rearrange the elements in each row such that the facets align vertically we get Figure 2.

	What	Where	When	Who
WHAT (LCSH)	X	X	X	X
WHERE (Place name list)	X	X	X	-
WHEN (Time period dir.)	X	X	X	-
WHO (biographical dict.)	X	X	X	X

Figure 2: Realignment of Figure 1 by facet, with examples of links. Vertical mappings extend semantic links to new vocabularies. Horizontal links (elements within a single records) provide additional context.

Present practice in linked data appears dominated by same-facet links within same-facet genres, e.g. mapping the names in one place name list to names in another place name list, or links among subject access vocabularies (UDC and LCSH). But to improve resource discovery we could benefit from same facet links between vocabularies in different facet genres, represented by vertical arrows in Figure 2. And if we are to understand how anything fits into its context we must also have links between different facets with the same facet genre, represented by horizontal arrows. For example the LCSH subject heading “Lighthouses” in a library catalog could quite easily link to the

U.S. National Geo-Intelligence Agency's geographical feature type "Lthse (Lighthouse)" in a place name list. The catalogue subject heading will lead to literature about lighthouses and the place name list will lead to and from the locations of actual lighthouses. I believe that Figure 2 indicates a very promising program.

BIBLIOGRAPHIC CITATIONS AND BIBLIOMETRICS

Bibliographic citations are a form of link that we could also examine. In this case, in addition to the immediate and direct usefulness of such links, there is now a veritable industry of bibliometric analysies that has a acquired a powerful but questionable role in the evaluation of individuals, academic departments, and universities -- and even in the allocation of resources. It is questionable because the assumptions underlying the use of the data are not entirely satisfactory (Gingras, 2013).

A FIELD THAT IS NOT PERFECT

The kind of work reflected in this symposium can be seen as practical and very useful. But it not perfect because when we look carefully we see difficulties in theory and in details:

- Links translate between one vocabulary and another, but a little something tends to be lost in translation;
- Some very useful kinds of relationship are expressed by links, but they are only a very small subset of the kinds of links that exist;
- The topics linked are primarily static when our world is dynamic;
- The commonest link relationship--<sameAs>--is very powerful, but what are linked are not strictly the same;
- Bibliometric links and bibliometric analyses are popular, powerful and influential, but we know that the assumptions are questionable and the conclusions are biased.

At first sight, it is disappointing, from a scientific point of view, to accept that our best achievements are useful but fundamentally imperfect. But I suggest that, to the contrary, it is a good conclusion. It is good because it is honest and it is good for a practical reason. There are a number of well-established ways to study human behavior, such as Economics, Linguistics, and Political Science. As Paul Ghils has pointed out (Ghils (2103), it is a basic characteristic of each of these fields that each contains some highly rational (formal, logical) techniques. In Economics is a good example. Economics includes microeconomics, a most powerful and respectable way to calculate the optimal decision in varying circumstances. But we also know that although individuals are more or less reasonable. They try to make sense of their financial situation, they are not totally rational. Humans general accept satisfactory solutions rather that insist on the strictly optimal outcome. Similarly, with macroeconomic projections of inflation, gross domestic product, and so on. The calculations are powerful and unquestionably useful, yet we know that they are based on data collection that is not perfect, that assumptions have been made. So, on this analysis, our field, Documentation, as Otlet called it, or Information Science as others do now, is just another field that analyzes and seeks to help an aspect of human activity. I say "just another", because for a long time there has been anxiety and self-consciousness about what kind of a field it is: Is it a science? Is it a discipline? Is it interdisciplinary? Is it a meta-science? -- and so on. Preparing for this Seminar has led me to the conclusion that it is the weaknesses of what we do that makes "just another" field, comparable to others concerned with human activity. For me, to be accepted as a sibling of Economics, Linguistics, and Political Science is a good outcome.

SUMMARY

1. Learning, knowing, understanding constitute how we live, so Documentation (by whatever name) is a form of cultural engagement. (Buckland 2012).
2. Our systems are full of links of many kinds, including subject indexes, syndetic structures, search term recommender services, as well as "linked data" in sense of Linked Open Data.

3. There is a tension between logic and language. Natural language is endlessly expressive and adaptive. Logic, however, is not. Natural language is capable of reflecting well the uncertainties, ambiguities, and dynamic change with which we make sense
4. Probabilistic methods can be useful and cost-effective in this complex and unstable world.
5. It is a challenge to embrace the full expressive power of language and acknowledge the cultural complexity of our environment.

The Seminar theme is rich at two levels: Exciting new tools for combining classification, authority control, and linking for improved research discovery; but also an invitation to examine more deeply the fundamental challenge of using formal rational tools in contexts that resist them: phenomena, language, culture, and knowledge.

Our seminar agenda is to explore how classifications, vocabulary control and links can expand resource discovery. The prospect of combining methods developed for World Wide Web with our own existing techniques is very promising. There is, however, a tension between standardized relationships, symbolized by Otlet's modernistic universalism and the Semantic Web, and the particular, subjective situations in which individuals try to make sense, reflected in Ludwik Fleck's emphasis on the role of local, temporary cultural contexts. This is important because Documentation exists to promote knowing and is, therefore, a form of cultural engagement.

If my analysis is correct, we have a dilemma: Our field has very powerful, very useful tools *and yet* these tools, if taken too far, reflect reality decreasingly. We see the same in Economics where microeconomic analysis is extremely powerful in determining optimal decisions and yet humans, when observed, do not in reality make sense with the hyper-rationality assumed in microeconomic analysis.

The challenge faced in this Seminar is to accept the full expressive power of language, to acknowledge the cultural complexity of our environment, *and* to find ways to use and improve the tools available to us. We can take pride in what we, Otlet's heirs, have achieved, but at the same time we need to take account of the insights of Ludwik Fleck as well as the vision of Paul Otlet.

Acknowledgments

References

American Library Association. Subcommittee on Subject Relationships/Reference Structures. (1997). *Final report to the ALCTS/CCS Subject Analysis Committee.* Available at: http://www.ala.org/alcts/mgrps/camms/cmtes/ats-ccssac/srrsreport.

Buckland, M. (2007). Naming in the library: marks, meaning and machines. In: *Nominalization, nomination and naming in texts*. Edited by C. Todenhagen, W. Thiele. Tübingen, Germany: Stauffenburg Verlag, pp. 249-260. Available at http://people.ischool.berkeley.edu/~buckland/naminglib.pdf.

Buckland, M. (2007). On the cultural and intellectual context of European documentation in the early twentieth century. In: *European modernism and the information society: informing the present, understanding the past*. Edited by W. B. Rayward. Aldershot, UK: Ashgate, pp 45-57.

Buckland, M. (2011). Obsolescence in subject description. *Journal of Documentation*, 68 (2), pp. 154-161.

Buckland, M. (2012a). Interrogating spatial analogies relating to knowledge organization: Paul Otlet and others. *Library Trends*, 61 (2), pp. 271-285.

Buckland, M. (2012b). What kind of science can Information Science be? *Journal of the American Society for information Science and Technology*, 63 (1), pp. 1-7.

Buckland, M. et al. (1999). Mapping entry vocabulary to unfamiliar metadata vocabularies. *D-Lib Magazine* 5 (1). Available at: http://www.dlib.org/dlib/january99/buckland/01buckland.html.

Buckland, M. et al. (2000). Variation by subdomain in indexes to knowledge organization systems. In: *Dynamism and stability in knowledge organization: Proceedings of the Sixth International*

ISKO Conference, 10-13 July 2000, Toronto, Canada. Edited by C. Beghtol, L. C. Howarth, N. J. Williamson. Würzburg, Germany: Ergon Verlag, pp. 48-53.

Buckland, M., A. Chen, F. C. Gey, R. R. Larson, R. Mostern & V. Petras. (2007). Geographic Search: Catalogs, Gazetteers, and Maps. *College & Research Libraries* 68, no. 5 (Sept 2007): 376-387. http://crl.acrl.org/content/68/5/376.full.pdf+html

Buckland, M.; Ramos, M. R. (2010). Events as a structuring device in biographical mark-up and metadata. *Bulletin of the American Society for Information Science and Technology* 36 (2), pp. 26-29. Available at: http://www.asis.org/Bulletin/Dec-09/Bulletin_DecJan10_Final.pdf

Day, R. E. (2013). "The data—it is me!" (les données—c'est moi!" In: B. C. Cronin & C. R. Sugimoto, eds. *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact*. (pp 676-84). Cambridge, MA: MIT Press.

Day, R. E. (2014). *Indexing it all: The subject in the age of documentation, information, and data*. Cambridge, MA: MIT Press.

Deutsche National Bibliothek. (2011) *AgRelOn, an Agent Relationship Ontology*. Available at: http://d-nb.info/standards/elementset/agrelon.owl#RelatedAgent.

Dewey, M. (1899). Dewey's *Decimal Classification and relativ index for libraries, clippings, notes, etc.* Boston: Library Bureau.

Fleck, L. (1935/1979). *Entstehung und Entwicklung einer wissenschaftlichen Tatsache: Einführung in die Lehre vom Denkstil und Denkkollectiv*. Basel: Schwabe, 1935. English ed.: *Genesis and development of a scientific fact*. Chicago: University of Chicago Press, 1979.

Functional Requirements for Bibliographic Records (*FRBR),* 1997. http://www.ifla.org/VII/s13/frbr/

Functional Requirements for Subject Authority Data (*FRSAD*), 2010. www.ifla.org/node/5849

Ghils, P. 2013. The essential tension: rational and reasonable in science and philosophy. *Transdisciplinary Journal of Engineering & Science* 4 (Dec 2013): 40-56.

Gingras, Yves. (2014). *Les derives de l'évaluation de la recherché: du non usage de la bibliométrie*. Paris: Raisons d'Agir Editions.

Hayes, P. (2011). On being the same as. Keynote address at the International *UDC Seminar 2011: Classification & Ontology: Formal Approaches and Access to Knowledge, The Hague (Netherlands) 19-20 September. Presentation a*vailable at: http://www.udcds.com/seminar/2011/media/slides/UDCSeminar2011_PatrickHayes.pdf

Hill, L. L. (2006). *Georeferencing : the geographic associations of information*. Cambridge, Mass. : MIT Press.

Konrad, A. (2007). *On inquiry: human concept formation and construction of meaning through library and information science intermediation*. Dissertation, University of California, Berkeley. Available at: http://escholarship.org/uc/item/1s76b6hp#page-1.

Larson, R. R. (1991). Classification clustering, probabilistic information retrieval, and the online catalog. *Library Quarter*ly, 61 (2), pp.133-173..

Larson, R. R. (1992). Experiments in automatic Library of Congress Classification. *Journal of the American Society for Information Science,* 43 (2), pp.130-148.

McKenzie, D. F. 1986. *Bibliography and the sociology of the text*. London: British Library.

OAEI 2014: Instance Matching Track (2014). Available at: http://islab.di.unimi.it/im_oaei_2014/index.html.

Otlet, P. (1934). *Traité de documentation*. Brussels: Editiones Mundaneum.

Otlet, P. (1935). *Monde: Essai d'Universalisme*. Brussels: Editiones Mundaneum.

Otlet, P. (1990). Something about bibliography. [Translation of "Un peu de bibliographie." Palais 1891/92: 254-271]. In: Otlet, P. 1990. *International Organisation and Dissemination*. Amsterdam: Elsevier, pp.11-24.

Petras, V. (2006). *Translating dialects in search: mapping between specialized languages of discourse and documentary languages*. Dissertation, University of California, Berkeley. Available at: http://www.sims.berkeley.edu/~vivienp/diss/vpetras-dissertation2006-official.pdf.

Rayward, W. B. (2013). From the index card to the World City: knowledge organization and visualization in the work and ideas of Paul Otlet. In: *Classification and visualization: interfaces to knowledge: International UDC Seminar 2015*. Edited by A. Slavic, A. Akdag Salah & S. Davies.

Würzburg: Ergon Verlag, pp. 1-42. Presentation available at: http://seminar.udcc.org/2013/programme.php.

Sady, W. (2012). Ludwik Fleck. In: *The Stanford Encyclopedia of Philosophy* (Summer 2012 Ed.). Edited by E. N. Zalta. Available at: http://plato.stanford.edu/archives/sum2012/entries/fleck/.

Standards for Networking Ancient Prosopographies: Data and Relations in Greco-Roman Names. (SNAP:DRGN). (2015). Available at: http://snapdrgn.net/.