

Preprint of Huvila, I. (2016). Some remarks on the possibility of extending the theory of polyrepresentation to the study of the relevance and qualities of 'things' represented in information collections. Farkaš, L.; Vican, D.; Tanackovič, S. F. & Ivanovič, M. D. (Eds.) *Ogledi o informacijskim znanostima - Zbornik radova u čast Tatjane Aparac-Jelušič, Filozofski fakultet u Osijeku & Sveučilište u Zadru*, , 187-203.

Some remarks on the possibility of extending the theory of polyrepresentation to the study of the relevance and qualities of 'things' represented in information collections

Isto Huvila

Department of ALM, Uppsala University

isto.huvila@abm.uu.se

### Abstract

This chapter explores some possible directions of future research on the contextual aspects of the phenomenon of polyrepresentation and the usability of the notion to understand the nature and scope of relevance of informative 'things' held at diverse collections and the nature of the collections. The eventual relevance of the proposed approach relates to bringing together contextual information from across the broad spectrum of general and specialised information collections and institutions including libraries, museums, archives and a broad range of public and commercial web services and thus contribute to understanding of the premises of the interoperability of information collections.

**Keywords:** polyrepresentation, information collections, scopes of relevance

### 1 Introduction

The polyrepresentation theory posits that when multiple different types of representations both in the cognitive space of the user and the information space consisting of information objects point to a particular set of information objects, it is probable that these objects are relevant (Ingwersen, 1996). The theory is based on three hypotheses. The first assumption is that the diverse representations in an information space and in the cognitive space of users can be matched. Secondly, it is assumed that more overlap between these representations pointing to one information object implies that the object is more likely to be 'relevant' (Ingwersen, 1994). Thirdly, it has been suggested that if the overlapping representations are functionally different, the likelihood of the relevance of information object is higher (Ingwersen and Järvelin, 2005, 206-209). The theory itself suggests of higher relevance potential of (basically) any information object with different overlapping cognitive and functional representations (Ingwersen and Järvelin, 2005,

351), but so far the focus of research on polyrepresentation has been on journal articles instead of other information objects. Besides focusing on particular types of information objects, polyrepresentation research has explored very little the specific characteristics of different types of 'representations' and their relation to the resulting relevance. The question of the diversity of information objects is especially focal with crowdsourced forms of information that differ considerably from the relatively homogeneous and well structured information objects like journal articles.

The aim of this chapter is to explore some possible directions of future research on the contextual aspects of the phenomenon of polyrepresentation and to discuss whether and how the patterns of the co-occurrence of the representations of information objects (or *referents*) are related to the contexts of the perceived relevance of their referents. The central question is whether the patterns of polyrepresentation (co-occurrence) might be indicative both of the 'amount' (degree of polyrepresentation) and the scope of relevance (the combination of systems/descriptors in which the polyrepresentation occurs) of the referents. Abbasi and Frommholz (2015) have shown that it is possible to use clustering to form partitions of representations and at least in principle to improve search performance by guiding users to focus on specific partitions instead of the total cognitive overlap. A question remains, however, whether the patterns of polyrepresentation can be used to characterise the scopes of relevance of naturalistic clusters of referents occurring in the wild and the contexts of their origin in order to understand the nature of the referents, the clusters and their contexts of origin and how, why and when they might be relevant for different users. The proposals discussed are (Q1) whether it is possible to detect polyrepresentation overlap between topics in knowledge organisation systems, (Q2) what kinds of referents express polyrepresentation (for example), (Q3) is it possible to identify *scopes of relevance*, and (Q4) what are their characteristics. In contrast to the earlier empathetically information retrieval oriented research in polyrepresentation theory the focus of the present study is on investigating the applicability of polyrepresentation theory to shed light on qualitative aspects of representational overlaps and asking whether the notion of polyrepresentation and the scopes of relevance are usable in eliciting and bringing together contextual information from various types of knowledge organisation systems. Eventually this type of an approach can be useful in bringing together contextual information from across the broad spectrum of general and specialised information collections and institutions including libraries, museums, archives and a broad range of public and commercial web services and thus contribute to the understanding of the premises of the interoperability, convergence and divergence of information collections. As a first step, instead of attempting to present a comprehensive empirical validation of the explored proposals, the present chapter discusses some of the premises for developing such studies in the context of social knowledge organisation (SKO) systems.

## 2 Polyrepresentation

The theory of polyrepresentation has been developed from the early 1990s onwards mainly by Ingwersen together with his colleagues. The theory is based on observations made on earlier studies of information needs and information retrieval on the redundancy of representations and that a combination of different types of representations of same information objects tend to produce better information retrieval results than when only single type of representations is used (Ingwersen, 1994). The theory stems from the cognitive viewpoint of information science and an assumed potential value of matching the variety of representations existing in information objects and in the cognitive space of individuals (Ingwersen, 1994). In practice, a combination of overlapping representations with different sources and

functions (e.g. author assigned article titles and indexer assigned metadata) would be a sign of greater relevance of an information object. The theory builds on the idea of cognitive overlaps and the significance of the cases when the number of overlaps increase (Larsen and Ingwersen, 2002; Ingwersen, 1994). Extending the earlier work, Lioma et al. (2010) proposed a formalisation of the theory using Subjective Logic to model the different representations of information needs as *beliefs* with degrees of *uncertainty*. The preliminary model and the practical applicability of the concepts of beliefs and uncertainty were been tested in practice by Lioma et al. (2012). Another line of research with the broad aim of operationalising polyrepresentation theory in the IR context relates to explorations of the relation of polyrepresentation and quantum logic -based IR (Zellhöfer and Schmitt, 2011).

The theory of polyrepresentation has been tested in and supported by several studies (e.g. Skov et al., 2008; Schaer et al., 2012). Pao (1994) investigated intersections of citations, indexer chosen terms and author created text. Skov et al. (2004; 2006; 2008) have tested the theory with Cystic Fibrosis test collection with positive results. The use of references (representations, or at least cues of behavioural practices) seems to improve retrieval performance whereas measures like the number of citations does not (Ingwersen, 2012b). Functionally different overlapping representations seemed to improve precision (Skov et al., 2006). The polyrepresentation seems to function, but only with a limited number (3-4) overlapping representations. Increasing the number of representations seem to decrease the precision of searches (Skov et al., 2004, 2006). The observation is supported also by (Lund et al., 2006). Recently, Abbasi and Frommholz have applied the theory in cluster-based browsing (Frommholz and Abbasi, 2014; Abbasi and Frommholz, 2015).

Although the polyrepresentation has been shown to work best in highly structured contexts, Larsen and Ingwersen (2005) have identified a continuum of polyrepresentation from structured to unstructured solutions. Skov et al. (2006) suggest, however, that natural language is too weakly structured to reveal and support the real sense of polyrepresentation.

Hjørland (2006) has made critical remarks on the polyrepresentation theory and especially on how it has been based on the cognitive viewpoint of information science with a consequent lack of emphasis on other theoretical underpinnings of information science and more practically oriented studies that discuss the topic of representation. Hjørland notes that the variation and origin of representations can be explained by other viewpoints than the cognitive one. In this respect the overlap may be designated as domain-specific, socio-cognitive, cultural or socio-cultural instead of cognitive. His proposal of a domain-oriented theory of polyrepresentation underlines the variability of representations in information space. Indexing research has demonstrated the variability of language used to describe information objects and that the rate of cross-indexer agreement tends to be notably low. In practice, however, in spite of the fundamental theoretical differences between the approaches suggested by Ingwersen et al. and Hjørland, the question of polyrepresentation and the theory is about overlap of representations in document space and user space and its implications.

### 3 Hypothesis

In spite of the suggestions of its potentially broader scope, the extents and possible contexts of the explanatory power of the polyrepresentation theory have been studied very little. Larsen, Ingwersen and Kekäläinen (2006) list five different types of representations that may overlap in journal articles. The differences depend on cognitive factors,

function and style of the representations. Ingwersen and Järvelin (2005, 342) suggest that all media types have their own polyrepresentative models and give music as an example. Secondly, they suggest that the representations can refer to document contents beyond subject matter (Ingwersen and Järvelin, 2005, 351). Representations can be in form of colour, structural elements, links and metadata elements. Following the same line of reasoning, it may be reasonably assumed that the phenomenon of polyrepresentation does not need to be confined to overlapping representations of traditional types of media such as journal articles, images or music. It is therefore hypothesised that overlap of *any* functionally different representations residing in user space and information space may be assumed to be related to the potentially higher 'relevance' of *any* object of representation (referent, e.g. document, but also a concept or topic). This topical approach assumes that polyrepresentation theory might be used to make inferences on the relevance of 'things' on much broader level than only on the relevance of information objects *par excellence*. An apparent question remains how to identify different unorthodox types of references.

Besides the overlapping of 'what', a central aspect of polyrepresentation is the match between different types of representations. Larsen and Ingwersen (2005) discuss the matching of different types of representations in the context of the notion of polyrepresentation continuum and distinct different approaches to match representations in structured and unstructured contexts. The interest of polyrepresentation continuum approach is, however, focused on the techniques of determining matching and non-matching representations. As suggested in the introduction, we hypothesise that the overlap of representations might be indicative of the context of relevance of a particular referent. If a referent is represented in multiple functionally different contexts (e.g. certain documents, knowledge organisation systems), it may be assumed that the degree of polyrepresentation and the shared characteristics of the contexts of representation are indicative the scope of relevance of the referent. Abbasi and Frommholz (2015) have shown that certain kinds of 'fields' (or partitions of total cognitive overlap) can be constructed using clustering and published results that effectiveness of the utilisation of cluster-based polyrepresentation in helping users to browse a collection. From the IR perspective, the finding that search effectiveness can be improved using the method is interesting. From a more general information science perspective, it is also interesting whether these fields correspond with fields identified, named and discussed by individuals i.e. whether they correspond with social categories. We hypothesise that this is the case at least to a certain degree.

Another implication of the hypothesis is that the intersections of representations might be indicative of the scope (and context) of relevance of the referent. That is, using a simple example, if a user types a query 'nice dog', has uploaded a photo of a dog (a beagle) to a dog owners' discussion forum, has tagged a large number of URLs with 'dog' and 'beagle' on a particular social bookmarking service and in a social cataloging site has catalogued a book called "A story of a nice beagle", it is likely that a relevant result for the query is something related to beagles. The example may also be generalised. If a large functionally diverse group of representations of a referent (beagle) occur in contexts characterised by frequent use of adjective 'nice' (a representation of nicety), it may be assumed that beagles tend to be considered nice (in these particular contexts and contexts closely related to them). Following the reasoning of the polyrepresentation theory, beagles are more likely to be nice (or, considered to be, within a particular context and situation) if the functional variety of representations is broader.

We assume that the shared characteristics of represented topics determine the *scope of relevance* of the systems. In the example, the nicety of beagles forms one possible scope of relevance that connects the search system, discussion forum, bookmarking site and social catalogue. A shared scope of relevance makes particular systems a relevant context for

information interactions related to a particular context shared by the SKO systems. For instance, if related the representations of topics  $X_1..X_n$  are related in collections. A and B, the systems are likely to form a relevant context for interacting with information about the topics. Consequently, also the cognitive authority of the systems A and B is higher concerning the topics  $X_1..X_n$ .

The practical relevance of the presented hypotheses is that by mining publicly available document collections and knowledge repositories including both formal information collections held by information and heritage institutions and crowdsourcing based SKOs, it is possible to harvest large amounts of representational data and to match that with digital information from user contexts in order to construct advanced (poly)representational profiles of different referents and their contexts to benefit both management and retrieval of information. Similarly, the profiles can be used to elicit further information on the nature and relation of the collections and eventually how they overlap, converge and differ.

#### 4 Towards an empirical study

A possible approach to empirically test the validation of the hypothesis could be retrieve representations (i.e. objects) related to two sets of referents from social indexing services. The approach was chosen for the sake of simplicity of obtaining a corpus of useful data for the first experiment. In future studies, the investigation should be extended to other types of data. In this context representation should be defined, as in (Larsen and Ingwersen, 2005), as “any tangible and recordable entity that has occurred as the result of a transformation of knowledge structures of a cognitive agent”. In practice, an existing Wikipedia article, a set of retrieved objects in other services, are considered to be representations. A topic is expected to be “well-defined” (Miller, 1999) i.e. confidently distinguishable and identifiable. Such topics could be based, for instance, on LCSH terms, because of the relative comprehensiveness and universal scope of the thesaurus. Because LCSH is based on the literary warrant (Olson, 2000), LCSH terms can be expected to represent well-defined topics that exist with certainty.

The study was conducted by automatic matching of two systematic samples of topics and information objects from eight SKO systems. A PHP tool was developed to get a systematic 100th term sample of the LCSH data (version July 16, 2010). In case of the Answers.com data, a set of 733 questions was scraped from the Answers.com list of recent questions on March 2, 2010 (similar methods used in Vakkari, 2010). Both LCSH terms and Q&A data were cleaned of stop-words and non alpha-numeric characters. LCSH terms were also cleaned of subdivisions (subterms), because initial trials showed that the complexity of the terms (as also suggested in Lin et al., 2006) caused a dramatical decrease of recall rate in SKO systems. The topics that returned hits from multiple SKO systems were compiled to a list and a random sample of 100 items (referents) with 2-6 polyrepresentations was subjected to an exploratory qualitative analysis. The purpose of the analysis was to find indications of possible polyrepresentation rather than to make definite inferences of its quantitative characteristics. The items were classified in topical groups (e.g. Current affairs related topics tended to appear more often in Youtube and Technorati) using the constant comparative method (Glaser and Strauss, 1967) to reach an adequate level of validity and reliability for the purposes of the present study. The representations of each topic were checked visually (downloaded and read/watched) in order to detect apparent similarities in the representations of referent in different SKO systems.

When interpreting the findings it is important to take into account that the data gathering method has severe limitations. First, even if two different methods of constructing topics was used, the resulting bags of words are unlikely to match

with the the patterns of how topics are constructed and expressed in different SKO systems. Secondly, considering the differences between search functions available in individual SKO services and the automatic procedure of constructing the bags of words, the data set is at the best a very crude (and most likely a highly inexact) approximation of the topics represented in the SKO systems. However, in spite of the limitations, the dataset contains a diverse set of representations with varying origins on a broad sample of topics that makes it possible to look at their intersections, i.e. polyrepresentation of the represented and unrepresented topics, and to find evidence on the plausibility of the presented hypotheses. Investigation of the quantities of polyrepresentation, user studies and testing the effect on the performance of information retrieval is left intentionally for future research.

## 5 Results

(Q1) Is there polyrepresentation overlap between topics represented by LCSH, Answers.com and the SKO data?

The data suggests that the overlap of representations in different SKO systems varies (Table 1), but that all SKO systems overlap with each other to a degree. The performance of individual social web services seemed to vary between referents although Technorati seemed to be clearly the best and Twitter and Bibsonomy tend to be worst performing SKO systems. Because of the crude extraction and comparison method, the presented percentages and data have, however, only indicative value. A plausible conclusion is that there is polyrepresentative overlap but its degree and variation between different data sources requires further investigation.

	Bibsonomy	Delicious	Faves	Flickr	Technorati	Twitter	Youtube
LCSH	18,56% (636/3426)	26.82% (919/3426)	52.66% (1804/3426)	29,57% (1316/3426)	98.83% (3386/3426)	5.87% (201/3426)	72.03% (2468/3426)
Q&A	6.94% (50/720)	5.13% (37/720)	23.19% (167/720)	1.11% (8/720)	n/a	5.69% (41/720)	n/a

**Table 1:** Overlap of referents in SKO systems.

The qualitative analysis of the randomly selected 100 referents gave further indication of the reality of the phenomenon (polyrepresentation). For instance, referent 'Acacia' was represented by images and web pages relating to Acacia tree. In contrast to Yi and Chan (2009) study, the social KOSs do not map as well to LCSH. The difference can be explained by differences in the approaches of the studies. Yi and Chan compared individual tags to LCSH tree while the present study set out to identify similarity of referents represented by LCSH headings and the bulk content of the social KOSs. The rough method used in the present study may be refined in the future studies to incorporate more elaborate means to adapt to different spellings, wordings and synonyms. The downside of inclusiveness is, however, the increasing amount of noise created by completely or nearly homonymous, but unrelated, topics.

(Q2) What kinds of referents express polyrepresentation and in what services?

Part of the the good overall performance of Technorati can be explained rather confidently by the broad scope of blogs indexed in the service and the length of an average post compared to considerably shorter textual content available in other SKOs. Even if no quantitative comparison was performed on Wikipedia data, a qualitative analysis of randomly selected 100 topics showed potential for an even higher and more precise overlap.

The qualitative analysis revealed that polyrepresentation seems to be strongest with referents sharing following

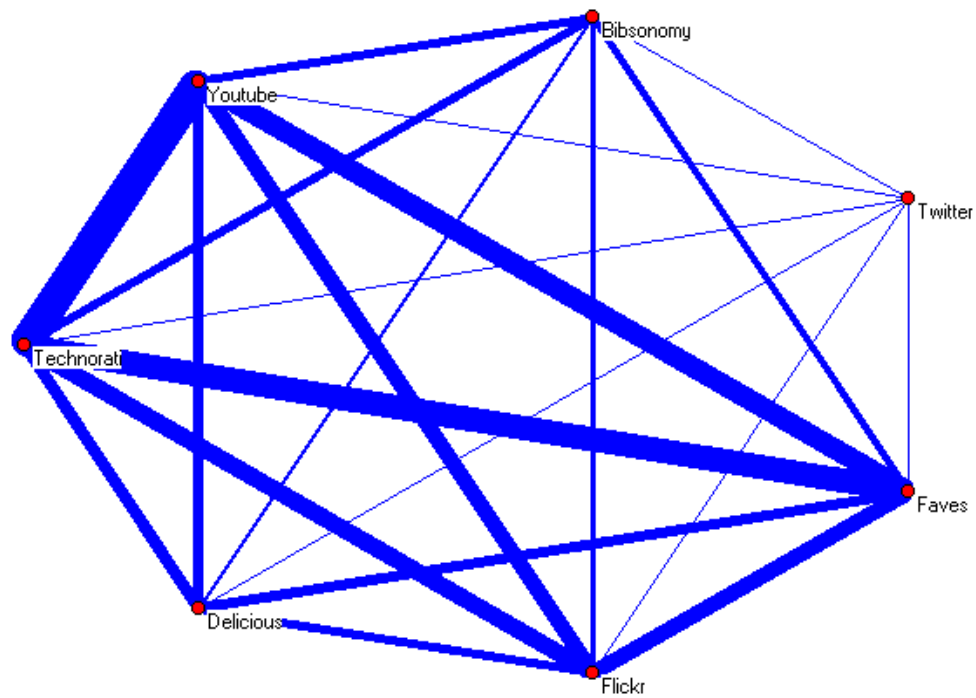
characteristics. Referent

1. Is comprehensible represented by a “name” formed by a low number of unambiguous words (e.g. Korea, Serb, softball)
2. Is of general interest i.e. belongs to “general knowledge” (e.g. sunflowers vs. agape or alloys)
3. Is concrete vs. abstract (e.g. animals, places vs. philosophical terms or emotions)
4. Is not specific to an epoch (due to the variation of expressing dates and epochs) (e.g. history vs. history of a specific region or epoch)
5. Has both current and long-term relevance (e.g. George Washington vs. new popular culture icons or forgotten authors)
6. Has both broader and personal frames of interest (e.g. country and place names and other topics that occur in both news items and personal notes and photographs)

In addition, the polyrepresentation showed clear variation between several SKOSs. Following the face-value logic, the referents with easily conceivable and concrete visual representations tended to have representations in image related social KOS. It seemed, however, that the ‘visuality’ of a referent might increase polyrepresentation also in other services. Both preliminary observations need to be tested further.

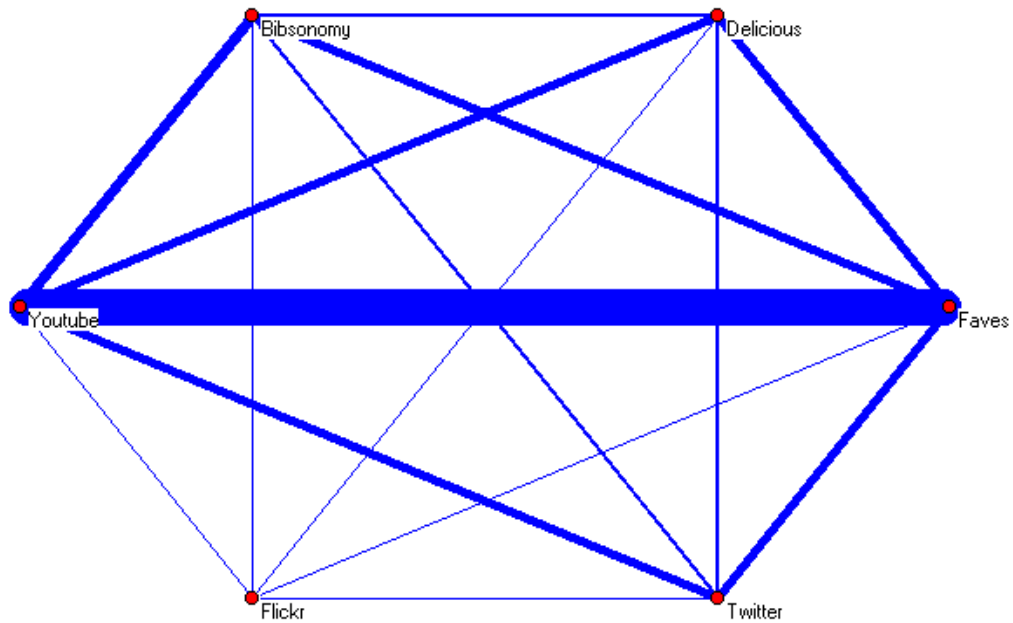
(Q3) Is it possible to identify *scopes of relevance* in the present data?

It seems to be so. The qualitative analysis revealed characteristics that can be used to explain the polyrepresentation of topics in SKO systems. In LCSH data, as indicated in the Figure 1, polyrepresentation was strongest between Technorati and Youtube. Both Faves and Flickr were strongly connected to the two. Twitter had clearly least polyrepresentations. With Answers.com data (without Technorati) the graphs is similar although Twitter has relatively more representations while Flickr has notably less.



**Figure 1:** A network graph of LCSH based polyrepresentations (line widths indicate the number of common representations) computed with Pajek (Kamada-Kawai algorithm, Kamada and Kawai, 1989).





**Figure 2:** A network graph of Answers.com based polyrepresentations (line widths indicate the number of common representations) computed with Pajek (Kamada-Kawai algorithm, Kamada and Kawai, 1989).

(Q4) What are their characteristics?

The data gives some indication of the existence of scopes of relevance. The overlaps may be partly explained by interlinking (especially Faves and Youtube) the similarities in the structural characteristics of the SKO systems. Social KOSs share an easily identifiable scope of interest related to *current affairs* that is strongest in Twitter and the blogosphere. In contrast, Wikipedia and Flickr share a *documentary* scope with inclination to form referents of longer term interest. Image and visual material related services share a *visual* scope of significance. Twitter, Technorati and Flickr share a *particularist* scope of significance. The contexts relate to individual cases of phenomenon while in other services the contexts are represented on a higher level. Flickr and Wikipedia form also a *geographic* scope of interest. The table presents examples of identified scopes of relevance.

If the topic consisted of only one (relatively) common word (e.g. abdomen, acacia) the polyrepresentation tended to extent to all SKO systems. However, even very specific terms were occasionally polyrepresented in the majority of the systems (e.g. Adenosylmethionine).

Scopes of relevance	SKO systems*	Example
Personal matters	Y,D,Fv,(Tw),(Fl),Te	abusive parents, anger in children
Professional settings	(B), Fv, De, Y, (T)	academic libraries, academic medical centers
How things work	Y, Te, F/De	Air bag restraint systems, APRS Telecommunication
News items	Tw, Y	Atlantic puffin
Geographical locations (small, specific)	Y, Fl, (De)	Anderson Ranch Dam Idaho, Ankobra River Ghana, Antioch Dunes California

**Table 2:** Homogeneous scopes of relevance (topic has similar representations in all SKO systems). Examples illustrate the referent, but do not necessarily polyrepresent in all or only in the indicated SKO systems. \* B=Bibsonomy, De=Delicious, F=Faves, Fl=Flickr, Te=Technorati, Tw=Twitter, Y=Youtube,. Systems in parenthesis refer to less frequent polyrepresentation.

In some cases topics get parallel meanings in some SKO systems. Angels in art returned angel and art related hits, but at the time of this study, in Twitter and Technorati, fashion related representations were more common in contrast to other services. The qualitative analysis of the randomly selected 100 referents indicated that it is possible to identify some tentative patterns of how different types of representations are polyrepresented in different SKO systems. Examples of the patterns are presented in Table 2.

## 6 Discussion

Apart from giving indications of that polyrepresentation occurs between LCSH and Answers.com, and SKO data, the analysis suggests that it seems to be possible to elicit and capture information about the scopes of the referents contained in the different systems and the contexts of the information behaviour of their users by identifying common characteristics of information objects. The third major observation is that polyrepresentation theory appears to be usable in making at least limited inferences on the 'relevance' (in the general sense of the term Cosijn and Ingwersen, 2000) of non-documentary (in the chapter, other types of objects than traditional structured documents such as journal articles and their metadata) information objects such as the ones represented by LCSH terms and the data harvested from the questions posted to Answers.com. What is the nature and scope of this 'relevance' requires further investigation similarly to how it is best explained either in the context of cognitive or other, for instance, domain-analytic or practice-oriented theories, or what are the practical applications of this observation.

The polyrepresentation theory is based on the assumption that a combination of overlapping representations with different sources and functions (e.g. author assigned article titles and indexer assigned metadata) would be a sign of greater relevance of an information object (Ingwersen, 1994). In case of SKO systems and non-documentary referents, it is apparent that the sources and functions of the systems differ from each other in many respects. Flickr is a photosharing service, Wikipedia an online encyclopaedia and Delicious a social bookmarking service. The variety of their users is equally easy to see. The variety was also confirmed in the qualitative analysis by the functional differences of the overlapping representations (i.e. when overlapping representations are used to refer to a group membership and

encyclopaedic representation of a topic).

Like Ingwersen and Järvelin (Ingwersen and Järvelin, 2005, 342) have suggested, the present study indicates that different 'media types' (i.e. in this study, SKO systems and their principal media forms) appear to have their own polyrepresentative models. Secondly, they suggest that the representations can refer to document contents beyond subject matter (Ingwersen and Järvelin, 2005, 351). The structural differences between the discussed SKO services are considerable and the different media types represented in different services did apparently skew the patterns of polyrepresentation. There are also a number of functional and style (Ingwersen and Järvelin, 2005, 208) related differences. Flickr tags have been found out to be mainly sourced by the authors of the photographs (Rafferty and Hilderley, 2007). The data from Technorati is from the texts of the blogs indexed by the service. It may be suggested that the differences in the representation of different topics in the indexing services owe to the *different (cognitive) interpretations* of same topic (Ingwersen and Järvelin, 2005, 277) and indexed representations of that particular topic. The same topic is represented in different SKO systems (i.e. services) with different domains and databases (Ingwersen and Järvelin, 2005, 277), but also different media types. In the context of multimedia information retrieval, Zellhöfer and Schmitt (Zellhöfer and Schmitt, 2011) have suggested that a quantum-logic based query model based on the exploitation of the non-simultaneous relevance of the different qualities of multimedia documents can be useful in improving retrieval outcomes. In this perspective, the suggestion of Ingwersen (Ingwersen, 2012a) that intermediate models of polyrepresentation may be more effective than total polyrepresentation seems to be applicable even in the present context of applying the polyrepresentation theory. Removing less-well performing representational combinations could improve the understanding of the contextual relevance and qualities of things both in the context of SKO systems and other information collections. The results of the present study do not provide means to make inferences about the precise characteristics of the polyrepresentative models of the investigated systems, but suggest that this could be a fruitful area of future research.

In the context of the polyrepresentation of web documents, Ingwersen and Järvelin (Ingwersen and Järvelin, 2005, 213) suggest that "the Web information providers are not following the principle of polyrepresentation" and that "the additional types of representations [other than outlinks and inlinks from and to web pages] are simply applied in pragmatic and often obscure ways". This may be the case with documents and on the level of conscious activity, but according to the present findings it seems that even though the web information providers are not explicitly following the principle, polyrepresentation exists on the web and functions on the level of topical representations even if the categorisation of individual documents and their relevance might be disturbed by noise caused by pragmatic and obscure representations.

In spite of the promising results, major reservations are necessary when the polyrepresentation of documents and referents are compared, because the phenomena are different albeit closely related. The results of the present study seem to indicate that a higher degree of polyrepresentation in different types of SKO systems is related to a higher relevance (or at least notability/popularity) of the referents. This seems to follow the face-value logic that more popular and in general sense relevant referents get a higher number of mentions and the theoretical assumptions of the polyrepresentation theory. In comparison to a generic trend analysis based on occurrence of keywords, the polyrepresentation based approach provides a theoretical framework to conceptualise the phenomenon of co-occurrence and to contextualise the co-occurring representations in the functional contexts of the SKO systems. It seems plausible to assume that similar patterns of polyrepresentation may be detected in formal collections of non-

documentary objects.

However, when interpreting the results, it is necessary to consider the several limitations of the empirical approach assumed in the present study. All findings should be considered as indications of potential future areas of study rather than as evidence of the reality and relevance of the preliminarily observed phenomena. More sophisticated content harvesting methods would be likely to provide more detailed and better contextualised sets of referents and representations in contrast to the rather simplistic approach assumed in the present study. The validity and reliability of contextual classifications was controlled by using constant comparative method, but in future studies, more robust methods of ensuring validity of the results could be used. In spite of its limitations, we argue that the assumed exploratory approach was adequate for the modest purposes of this first study.

## 7 Conclusions

The central question of this study was whether polyrepresentation (co-occurrence) might be indicative both of the 'amount' (degree of polyrepresentation) and the scope of relevance (indicated by the combination of systems/descriptors in which the polyrepresentation occurs) and in broader sense, of the qualities of naturalistic clusters of arbitrary referents occurring in the wild and of the contexts of their origin and whether the characterisation would be helpful in understanding the nature of the referents, the clusters and their contexts of origin and how, why and when they might be relevant for different users. On the basis of the present results it may be hypothesised that if the scope of relevance of the retrieved information objects can be approximated, at least in principle it can be used to select a set of better performing overlapping representations, and consequently, to better understand the 'nature' (or how it is perceived and expressed by individuals) of the referents for the purposes managing and searching relevant information. The overlapping representations form a context for information interactions with a higher likelihood of significance. This could be applied, at least in principle, to combinations of references from any types and combinations of referents. Considering the earlier reservations on the applicability of weakly structured references (e.g. Skov et al., 2006), it is apparent that there are several practical issues to solve. The present study identified tentative scopes of relevance such as general information, photographs, current affairs (Technorati) and technical resources (Delicious). These particular scopes might not be highly useful in practice but we assume that by developing the methods of identifying scopes in future studies, it might be possible to identify scopes with more direct practical relevance. It is also possible to speculate that the polyrepresentative patterns of things can further be used to understand other aspects (than their scopes of relevance) of their referents from their perceived (cognitive and other forms of) authority, context (of use) and role that would open up for broader questions on the informativeness of things held in diverse collections at institutions like libraries, archives and museums and beyond, and eventually contribute to the comprehensive understanding of the prospects of bringing together assets held by these institutions.

## References

Abbasi, Muhammad Kamran, and Ingo Frommholz. "Cluster-based polyrepresentation as science modelling approach for information retrieval." *Scientometrics* 102, 3: (2015) 2301–2322.

<http://dx.doi.org/10.1007/s11192-014-1478-1>.

Cosijn, Erica, and Peter Ingwersen. "Dimensions of relevance." *Information Processing and Management* 36, 4: (2000) 533–550.

Frommholz, Ingo, and Muhammad Kamran Abbasi. "On Clustering and Polyrepresentation." In *Lecture Notes in Computer Science*, edited by Maarten de Rijke, Tom Kenter, ArjenP. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann, Springer International Publishing, 2014, volume 8416, 618–623–.

Glaser, B. G., and A. L. Strauss. *The discovery of grounded theory: Strategies for qualitative research*. Hawthorne: Aldine, 1967.

Hjørland, Birger. *Lifeboat of Knowledge Organization*, Copenhagen, 2006, chapter Polyrepresentation. [http://www.db.dk/bh/lifeboat\\_ko/CONCEPTS/polyrepresentation.htm](http://www.db.dk/bh/lifeboat_ko/CONCEPTS/polyrepresentation.htm).

Ingwersen, P. "Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction." In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. New York: Springer, 1994, 101–110.

\_\_\_\_\_. *Scientometric Indicators and Webometrics – and the Polyrepresentation Principle*, volume 28 of *Sarada Ranganathan Endowment Lectures*. New Delhi; Bangalore: ESS Publications, 2012a.

Ingwersen, Peter. "Cognitive Perspectives Of Information Retrieval Interaction: Elements Of A Cognitive IR Theory." *Journal of Documentation* 52, 1: (1996) 3 – 50.

\_\_\_\_\_. "Citations and References As Keys to Relevance Ranking in Interactive IR." In *Proceedings of the 4th Information Interaction in Context Symposium*. New York, NY, USA: ACM, 2012b, IIIX '12, 1–1. <http://doi.acm.org/10.1145/2362724.2362726>.

Ingwersen, Peter, and Kalervo Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Berlin: Springer, 2005.

Kamada, T., and S. Kawai. "An algorithm for drawing general undirected graphs." *Information Processing Letters* 31, 1: (1989) 7–15.

Larsen, Birger, and Peter Ingwersen. "The boomerang effect: retrieving scientific documents via the network of references and citations." In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2002, 397–398.

\_\_\_\_\_. "Cognitive Overlaps along the Polyrepresentation Continuum." In *New Directions in Cognitive Information Retrieval*. Springer Netherlands, 2005, volume 19 of *The Information Retrieval Series*, 43–60. <http://www.springerlink.com/content/qn68220536xm31xt/>.

Larsen, Birger, Peter Ingwersen, and Jaana Kekäläinen. "The polyrepresentation continuum in IR." In *IIIx: Proceedings of the 1st international conference on Information interaction in context*. New York, NY, USA: ACM, 2006, 88–96.

Lin, X., J.E. Beaudoin, Y. Bui, and K. Desai. "Exploring characteristics of social classification." In *Proceedings 17th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research*. 2006, volume 17.

Lioma, Christina, Birger Larsen, and Peter Ingwersen. "Preliminary Experiments Using Subjective Logic for

- the Polyrepresentation of Information Needs.” In *Proceedings of the 4th Information Interaction in Context Symposium*. New York, NY, USA: ACM, 2012, IIX '12, 174–183.  
<http://doi.acm.org/10.1145/2362724.2362755>.
- Lioma, Christina, Birger Larsen, Hinrich Schuetze, and Peter Ingwersen. “A Subjective Logic Formalisation of the Principle of Polyrepresentation for Information Needs.” In *Proceedings of the Third Symposium on Information Interaction in Context*. New York, NY, USA: ACM, 2010, IIX '10, 125–134.  
<http://doi.acm.org/10.1145/1840784.1840804>.
- Lund, B., J.W. Schneider, and P. Ingwersen. “Impact of relevance intensity in test topics on IR performance in polyrepresentative exploratory search systems.” In *EESS 2006: Proceedings of the ACM SIGIR 2006 Workshop on Evaluating Exploratory Search Systems*. 2006, 42–46.
- Miller, George A. “On Knowing a Word.” *Annual Review of Psychology* 50, 1: (1999) 1–19.  
<http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.psych.50.1.1>.
- Olson, H.A. “Difference, culture and change: The untapped potential of LCSH.” *The Cataloging & Classification Quarterly* 29, 1/2: (2000) 53–71.
- Pao, Miranda Lee. “Relevance odds of retrieval overlaps from seven search fields.” *Information Processing & Management* 30, 3: (1994) 305 – 314. <http://www.sciencedirect.com/science/article/B6VC8-469V2FJ-3R/2/9b2ae54861abe25a6c5d05fe391069aa>.
- Rafferty, Pauline, and Rob Hilderley. “Flickr and Democratic Indexing: dialogic approaches to indexing.” *Aslib Proceedings* 59, 4/5: (2007) 397–410.
- Schaer, Philipp, Philipp Mayr, and Thomas Lücke. “Extending Term Suggestion with Author Names.” In *Lecture Notes in Computer Science*, edited by Panayiotis Zaphiris, George Buchanan, Edie Rasmussen, and Fernando Loizides, Springer Berlin Heidelberg, 2012, volume 7489, 317–322–.
- Skov, M., H. Pedersen, B. Larsen, and P. Ingwersen. “Testing the principle of polyrepresentation.” In *ACM SIGIR 2004 Workshop on Information Retrieval in Context*. 2004, 47–49.
- Skov, Mette, Birger Larsen, and Peter Ingwersen. “Inter and intra-document contexts applied in polyrepresentation.” In *IIX: Proceedings of the 1st international conference on Information interaction in context*. New York, NY, USA: ACM, 2006, 97–101.
- \_\_\_\_\_. “Inter and intra-document contexts applied in polyrepresentation for best match IR.” *Information Processing & Management* 44, 5: (2008) 1673 – 1683. <http://www.sciencedirect.com/science/article/B6VC8-4T083DM-1/2/0dc3bfaca401b03ba9131c6e1bce4a41>.
- Vakkari, Pertti. “How specific thesauri and a general thesaurus cover lay persons’ vocabularies concerning health, nutrition and social services.” In *Paradigms and conceptual systems in knowledge organization: Proceedings of the Eleventh International ISKO Conference*, edited by C. Gnoli, and F. Mazzocchi. Würzburg: Ergon, 2010, volume 12 of *Advances in Knowledge Organisation*, 299–307.
- Yi, Kwan, and Lois Mai Chan. “Linking folksonomy to Library of Congress subject headings: an exploratory study.” *Journal of Documentation* 65, 6: (2009) 872 – 900.
- Zellhöfer, David, and Ingo Schmitt. “A User Interaction Model Based on the Principle of Polyrepresentation.” In *Proceedings of the 4th Workshop on Workshop for Ph.D. Students in Information & Knowledge Management*. New York, NY, USA: ACM, 2011, PIKM '11, 3–10.

<http://doi.acm.org/10.1145/2065003.2065007>.

Zellhöfer, David, and Ingo Schmitt. "Approaching Multimedia Retrieval from a Polyrepresentative Perspective." In *Lecture Notes in Computer Science*, edited by Marcin Detyniecki, Peter Knees, Andreas Nürnberger, Markus Schedl, and Sebastian Stober, Berlin & Heidelberg: Springer, 2011, volume 6817, 46–60.