ART DISCOVERY GROUP CATALOGUE

FINAL GOAL OR STARTING POINT?

By Jan Simane

Lecture given at the 6th artlibraries.net meeting, Copenhagen, October 10–12, 2014

Looking back over the last two years, from the birth of the idea to the current state of development of an innovative, future-oriented model for art bibliography, based on the WorldCat architecture, we can certainly be happy with the result. We have achieved more than expected in terms of the number of participating libraries, coping with the technical and financial challenges as well as the high level of interest in the project, and far-reaching acknowledgment of the method in our professional community. Furthermore, the concept of our group catalogue has apparently piqued the curiosity of OCLC professionals and it could become a model for other disciplines. Finally, we have been and are still involved in the exciting experience of developing WorldCat's striking potential to serve primarily as a comprehensive information source, going far beyond the mere aggregation of holdings-based library catalogues. There is broad consent in the project team and among cooperating partners to identify and integrate relevant resources and data collections as well as new participating libraries and thus to enhance the quality and complexity of our discovery environment. It is not too bold to say that the 2010 art bibliography crisis has to a certain degree been overcome and that the Art Discovery Group Catalogue represents precisely what was predominantly demanded: a modern alternative to the outdated traditional bibliography format.

So, can we lean back and enjoy the pleasant 'mission-accomplished feeling'? Certainly not. However, I do not intend to list the many steps that still have to be taken and the hurdles that still have to be overcome in order to perfect what has been implemented. Rather, I would like to discuss some key issues of the new identity we all – as participating libraries – have to face after our catalogue data, and sometimes only single elements of it, has been merged and clustered into new bibliographic and cooperatively compiled units. We are experiencing something like an 'amalgamation' – at least in terms of display and access – of our metadata, a process that can blank out entries when additional information is missing and limit their significance to the secondary holding level. Of course, many of our libraries are integrated within networks and consortia where a similar workflow philosophy is the norm, first and foremost for economic reasons. Our Art Discovery Group Catalogue, however, has

other aims. Unlike in library consortia, the objective is not the consistency of the cataloguing rules or the most efficient streamlining of the cataloguing work but rather the compilation of relevant information originating from many different sources. This can be achieved in two ways: first, when missing records are added and the data pool of the Group Catalogue and consequently of the WorldCat itself is thereby enriched. Secondly, when further information derived from uploaded new data collections has been appended to already existing WorldCat records. The second process is obviously conducted with the help of specific algorithms and it should only be mentioned in passing that OCLC is investing a considerable amount of research resources into improving the methods for identifying, assembling and linking relevant and related data stored in the WorldCat records.[1] For the benefit of our Group Catalogue we very much hope that the successful policy of creating a multinational network of partners – a policy that has been practiced for over ten years in the artlibraries.net initiative – will increase the number of additional bibliographic records and lead to the first aforementioned effect. But we also have to realize that the entering of unique material will be rather the exception and that in most cases the aforementioned procedure of merging and clustering will be the norm. In other words, algorithms and uploading processes categorize and condition our catalogue entries in terms of both exposure and indexing with the result that the significance of any single bibliographic record will be evaluated according to its information input. In this respect most of us are engaged in a new experience with the WorldCat. As long as we do not catalogue within the WorldCat itself – and this situation will presumably not change in the near future – we will upload to and accumulate our locally produced records in the WorldCat which are to a certain extent redundant, to a certain extent new and to certain extent complementary with already existing records due to the additional information they contain. In this context it does not matter that the related algorithm-based procedures are not always able to process the data in the desired manner. It is however important to recognize that our records start to interact with other records when they are uploaded to the WorldCat. This is the fundamental difference between union catalogues and federated search environments like artlibraries.net. In other words, the ontological principles of our catalogue data change in the moment they are uploaded to the WorldCat.

Why is this so important? Or is it important at all? It is undoubtedly important when we believe in the vision of building up an alternative model to traditional bibliographies and cooperating in the – so to speak – protected domain of a group catalogue in order to develop tools and search criteria for a discipline-

[1] Cf. Gatenby, Janifer ed. al., GLIMIR: Manifestation and Content Clustering within WorldCat, in: Code4lib journal, 17, 2012 (http://journal.code4lib.org/articles/6812).

specific discovery experience. Of course, this is a rather theoretical approach to the current state of project development. But visions need a theoretical basis, otherwise they are not visions but pragmatic action plans. However, our vision should be realistic. So what we need is to gain a better understanding of the new reality of our data. In this respect we should recall the central goal of the project: we want to promote the discovery experience and the new tool should first and foremost support searches for documents and sources that are unknown to the user. Moreover, we want to emphasize that in view of the high number of records and the inevitable long results lists the filtering and selecting tools must be suitable for discipline-specific requirements. They must be reliable and of constant quality, and this in turn very much depends on the quality, or better on the nature, of the data itself. This sounds like a matter of course, but when we look at our bibliographic records and how they interact in this initial phase of coming together we have to admit that the reality can sometimes be disappointing. Before looking at some examples it is first important to understand the core principles of navigation as well as users' expectations as far as discovery systems are concerned.

From the user's point of view, the new grouping of art libraries catalogues into one discovery environment corresponds to a "googlisation" of the query + response process. It is not necessary to explain this step in detail as everybody here in the audience is familiar with this experience. Characteristic features are the default one-search field, the relevance-ranked results display and the facets found in other discovery systems rather than in generic search engines. Relevance ranking algorithms and facets both serve to manage the high number of results. OCLC has configured the related tools in a standardized form. As we learned in several discussions with OCLC representatives, the configuration can be modified to meet our requirements particularly when facets are concerned. However, for relevance ranking as well as for facets the quality of the underlying metadata is decisive. The relevance ranking algorithm in the WorldCat is built, as is true for most bibliographic databases, upon the statistical analysis of the frequency of key terms and their position in the record scheme. Put simply, the more often the search term appears in a record, in particular in the preferred author and title fields, the higher its relevance ranking. But this can happen by pure chance. Additional information like tables of content, abstracts, cover texts and the like can repeat the term in question many times, in contrast to other documents where such additional sources are missing although the relevance may be the same or even higher. Subject descriptors play an important role in this context. For very good reasons that are perfectly in line with our requirements, OCLC is clustering subject terms in merged records, independent from their origin and language. This means that the number of potential access points to one and the same document is growing. But at the same time the amount of

redundant information is also growing. Thus, the repetition of the same term, though it may be a consequence of the automatized clustering process, may rank this item higher even when the relevance of the title is not necessarily on a corresponding level.

There is a lot of inconsistency in this field, but this is hardly surprising if we consider how the WorldCat is aggregating the enormous quantity of bibliographic data. The subsequent problems, however, cannot be ignored. As we have seen, subject descriptors in particular have a strong influence on both the ranking and the faceting of an item. What is surprising is that apparently not all libraries uploading data to the WorldCat are also exposing and indexing their subject terms although they exist and are searchable in local catalogues, for instance at the INHA, the most important French library in our circle. Thus, many French subject terms are not effective for searches in the Art Discovery Group Catalogue. The same is true for some of our German partners such as the Berlin or Heidelberg libraries. We can find many examples of records which are very well described with subject terms at local level but these terms are missing in the WorldCat correspondents. Of course, this is the result of complex match and merge procedures and not all WorldCat libraries have the same philosophy in mind as our group: to serve first and foremost as a bibliographic source and to enhance the discovery experience as efficiently as possible. To achieve this goal we need appropriate tools to deal with the long results lists and these tools are – as mentioned before – currently based prominently on relevance ranking and faceting methods. We have learned that our data is not ideally organized for this challenge, in particular in the field of content description. We have to deal with both imbalance and inconsistency, phenomena found in all library catalogues, no doubt. However, in view of the growing number of comprehensive discovery environments the importance of standardizing metadata has been emphasized many times, in particular when library catalogues are integrated with other resources such as repositories, archives and the like into one-search-systems. In an OCLC research report published in 2011 Leah Prescott, former digital project coordinator at the Getty Research Institute, demonstrated these essentials and the related problems also in the challenging conversion of library, museum and archive collections with single-search access.[2] The single-search standard, closely interwoven with the default one search-field, is another, already mentioned phenomenon we are experiencing in the Art Discovery Catalogue being applied for the first time to a high number of art library catalogues and to all categories of additional information. Prescott's simple-sounding remark that "the quality and quantity of metadata affect the quality of the single search

[2] Prescott. Leah, Single Search: The Quest for the Holy Grail, OCLC Research 2011. http://www.oclc.org/research/publications/library/2011/2011-17.pdf. (September 2014)

experience" is attested by the consequent conclusion that "adherence to standards is essential, as integrating metadata is one of the most challenging aspects of a single search implementation."[3] This challenge is also valid for our Art Discovery Catalogue since we are integrating metadata and offering established tools and features we are used to applying in other discovery environments. All in all we have three relatively new issues that can be misleading when the current reality of our metadata is being blinded out: the single-search default (certainly the most preferred access method) will suggest that it is possible with one step to discover all the resources united in the aggregator. The relevance-ranked hit lists will suggest that the results be sorted according to the user's interest. And the facets will suggest that users can filter highly specific subsets of the hits. We can hardly satisfy these expectations with our catalogue data, not to mention all the additional sources which we are so happy about but whose metadata may be far removed from any bibliographic standard. My intention is anything but fatalistic. It is rather an attempt to gain a better understanding of what is happening with us and our catalogues after having made this substantial step from the passive existence of our data in the federated search network of artlibraries.net to the more active role it obtains or *can* obtain when its interrelatedness is more apparent. At the moment we must accept that there is still a gap between the potential of the discovery infrastructure on the one hand and the data we are providing for new services on the other.

Are these the typical considerations of a librarian to which users are generally indifferent? There is not much reason to believe so; in fact, it seems to be to the contrary. In 2009 an OCLC project team published a survey entitled "Online Catalogs: What Users and Librarians Want"[4]. Undergraduates, scholars and casual users were interviewed about their experiences with and requirements for searches in the WorldCat.org Database. It is interesting in our context that the ranking of search results according to relevance is generally acknowledged and taken seriously as a standard not only in generic search environments such as Google but also in modern library catalogues. What users expect is a reliable method of finding the most relevant titles that exactly match their needs. We have to admit that there are a lot of shortcomings in how relevance is defined in the WorldCat when the underlying records are as inconsistent as ours and all the others of course. The repeatedly shown parallels between library databases and generic search engines cannot hide the fact that these systems use much more differentiated and complex algorithms

[3] Ibd. p. 18.
[4] Calhoun, Karen ed al., Online Catalogs: What Users and Librarians Want, an OCLC Report, 2009 (www.oclc.org/reports/onlinecatalogs/fullreport.pdf) (September 2014)

for relevance ranking.[5] Thus, many users are apparently 'spoiled' by the quality of the relevance ranking algorithms of search engines and they expect a similar efficiency in library catalogues.

Looking at the facets the situation is different but no less difficult. The aforementioned OCLC survey of 2009 was addressed not only to users of the WorldCat but also to librarians. Not surprisingly the librarians saw the importance and need to enhance data quality in terms of accuracy and structure because "end users benefit from this structured data, for example when refining their searches with facets; however, end users tend to be unaware of the data quality requirements that support the functionality they rely on for precise, consistent results."[6] In principle, the traditional format of our catalogue entries is very suited to the facet system. Facets are 'access points' to metadata content and when these access points are clearly defined and uniformly filled with standardized categories of information the faceting of results can be constantly reliable and constructive.[7] But we have to face two problems: the number of distinct identifiers in a standard catalogue entry, such as year of publication, author etc., is limited and in the most important field, the content description with subject terms, the records are very heterogeneous and inconsistent. According to a study carried out in 2010 only 46% of the WorldCat records had subject terms at all and only 14% had DDC classes.[8] Moreover, as we have seen before, the high degree of redundancy in subject descriptions reduces even further the usefulness of these indicators for both faceting and relevance ranking. To repeat it once again: it is the interaction of the data, and not its mere aggregation, that is so decisive for the discovery experience in an environment like the WorldCat. The best algorithms can only cluster and activate the data and identifiers they can understand and extract from the records. Of course, this is a rather simple insight, and it serves to mention only parenthetically that the vision of a linked data world and all the related projects, such as the recent and ongoing Cornell-Harvard-Stanford initiative LD4L (Linked Data for Libraries), for instance, is based on this plain logic: the translation of relevant information into a standardized and preferably uniform encoding system in order to enable data interoperability in the semantic web.

As we come to its end, we may wonder what the intention of this presentation was. We can certainly exclude the requirement of the near-term harmoniza-

[5] Cf. Kinstler, Till, Making Search Work for the Library User, in: Catalogue 2.0: the Future of the Library Catalogue, ed. Sally Chambers, London 2013, pp. 25-26.
[6] Calhoun (n. 4), p. 40.
[7] Kinstler (n.5), p. 30.
[8] Cf. Implications of MARC Tag Usage on Library Metadata Practices, Karen Smith-Yoshimura ed al., 2010, p.19-20. www.oclc.org/research/publications/library/2010/2010-06.pdf.

tion of cataloguing conventions and practices. Such an expectation is completely utopian. Although I am convinced that in the foreseeable future many if not most libraries worldwide will catalogue directly into the WorldCat, it is not realistic to expect far-reaching uniformity in cataloguing praxis. All our experiences with union catalogues teach us the opposite. Moreover, it is obvious that in light of the aggregation of a high volume of harvested data from repositories, publishers' domains, archives and the like, with propriety metadata it is not very convincing to insist on uniform cataloguing standards for libraries. On the other hand the results would be more accurate, relevant and easier to process if the filtering principles of the discovery tool were more efficiently supported by corresponding data types. In this respect the ADGC is indeed a starting point. The bibliographic data that has been and still will be aggregated in this new environment does not interact perfectly because it has been produced independently and in many cases according to different search and retrieval purposes. We are now becoming aware that this data and the corresponding cataloguing traditions have to face the new ontology of the discovery domain where – to put it simply – the usefulness of our bibliographic records can rank rather high or rather low. It is difficult to predict whether this insight will change our cataloguing work or not. At present, I do not expect much. However, our initiative was prompted by the vision – here we are again – of building an innovative, efficient and appropriate tool for discipline-related information mainly based on catalogue data. Thus, the aggregation of our catalogues without any modification can hardly be the final goal. Already, an awareness of the coaction of our data and how it interacts with that of our partners could change things. We could try for example to provide bibliographic records with complementary rather than redundant information. This would be a big step forward. And of course, we are not alone. There is a lot of development and research at higher levels which will certainly increasingly relativize some of the shortcomings of our own data, but our data will always play an active role in this system of interaction and interrelatedness. To return to the beginning of my presentation: yes, we can be happy with the balance of the project in its current state, but the challenge to go further already awaits us.