Towards a Taxonomy for Web Observatories

Ian Brown
Web Science Institute
University of Southampton
Southampton, SO17 1BJ, UK
+44 (0)23 8059 5000
icb1g12@soton.ac.uk

Wendy Hall
Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, UK
+44 (0)23 8059 5000
wh@soton.ac.uk

Lisa Harris
Web Science Institute
University of Southampton
Southampton, SO17 1BJ, UK
+44 (0)23 8059 5000
l.j.harris@soton.ac.uk

ABSTRACT
In this paper, we propose an initial structure to support a taxonomy for Web Observatories (WO). The work is based on a small sample of cases drawn from the work of the Web Science Trust and the Web Science Institute and reflects aspects of academic, business and government Observatories. Whilst this is early work it is hoped, by drawing broad brushstrokes at the edges of different types of Observatory, that future work based on a more systematic review will refine this model and hence refine our understanding of the nature of Observatories. We also seek here to enhance a faceted classification scheme (which is thought to be weak in the area of visualisation) through the use of simplified concept maps.

Categories and Subject Descriptors
H.5.3 [**Group and Organisation interfaces**]

General Terms
Design, Standardization, Theory.

Keywords
Web Science, Web Observatory, Taxonomy, Observatory models.

1. INTRODUCTION
In the analysis of types of entities seen "in the wild" (natural or technological) it is often helpful to group/cluster the features, behaviours, structures and other phenomena according to classification schemes which can help in generating knowledge/insight about these entities. Kwasnick [1] asserts the linkage between classification and knowledge and describes several alternative structures for classifications/taxonomies such as Hierarchies, Trees, Paradigms and Facets. Spiteri [2] offers a selection process for a classification scheme which we have adopted. The definition of Web Observatories (beyond being a repository of data on the Web) is still developing and in these cases Spiteri argues against the use of hierarchies/trees. We also considered the multiple orthogonal features of WO's and the lack of automatic inheritance or transitive relationships between these features and thus we have elected to employ a faceted approach.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.

ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579212

Morshead [3] specifies that a Taxonomy requires both structure and a conceptual model for construction and herein perhaps lies the main challenge for a Taxonomy of Web Observatories (WO). Whilst the physical structures and data flows may be similar (or only trivially dissimilar) across differing WO implementations, the set of overarching concepts could (unhelpfully) reduce to no more than: (Data In, Process, Data Out) *unless* different perspectives are applied. Using a more refined analysis of the content and use innovation (the WHAT) and also the Actors (the WHO) and objectives (the WHY) rather than focusing only on the technical implementation details (the HOW) richer distinctions may be possible. We argue that physical implementation is less relevant according to Spiteri's test of faceted classifications in relational to other factors (though clearly not to the implementors of Observatories themselves).

A WO shares a number of characteristics within the family of data repositories and information systems (data warehouses, search engines, Big Data systems) and whilst business classifications of these more general systems have been attempted by groups such Gartner and IDC there is currently no such formal classification available for Web Observatories. This applies not only in terms of each WO as a standalone service, but also in relation to multiple interoperating WO's forming part of a global *World-Wide* Web Observatory (W^3O) which might exhibit additional synergistic or emergent properties. We have discussed the broad nature of these similarities/differences in earlier work [4] and in this paper we focus on generating a faceted taxonomy based on an analysis both of existing Observatories and future Observatory specifications.

2. BENEFITS
As this is still early work (as much work on Observatories is) our intention in moving towards a taxonomy is not to prematurely "lock down" the structure of Observatories but to enable three potential benefits: (1) To sketch out a "vocabulary" of elements giving a basis for designers, developers, and users/researchers to consider what is possible and what may be missing from existing systems and thus underpin potential gap analysis and design processes. (2) To provide a structure to consider how each node may interoperate with other nodes in a network of Observatories and thus underpin a framework for interoperability. (3) To highlight sub-types of Observatories (potentially with different research/analytical objectives) so that potential tensions around design and operation can be surfaced. Examples include the tension between open and closed content and commercial or not-for-profit operations, which, by definition, must (de) prioritise certain affordances/features.

The objectives of different groups using the Web in government, business and academia are not wholly aligned (i.e. the desire for

more or less anonymity, more or fewer pay-walls etc.) and we suggest that as part of the Web eco-system similar groupings of Observatories are likely to exist/develop. Categorising the drivers and intentions of certain Observatories may assist in harmonising interoperation between them.

3. METHOD

The methodology is based on Spiteri's simplified process for faceted analysis (from earlier work by Ranganathan [5] and Vickery [6]) with additional material from Denton [7]. It comprises the selection of an appropriate taxonomic structure (in this case facets) and then a step-wise process of generating facets, terms and finally items as part of a faceted classification scheme (FCS). Facets have been chosen here over hierarchies/trees or paradigms according to Kwasnick's criteria due to need for flexibility (*hospitality* as she calls it) in incorporating new facets/aspects in future models. Our simplified approach follows a four-step process comprising domain collection, entity listing, facet creation and facet arrangement. An iterative refinement of facets is considered as part of the reflection and testing process. Spiteri provides seven criteria for judging the choice of facets, which Kwasnick highlights as one of the key challenges of this approach along with a lack of ability to express relationships between facets and the visualization of facets. Our arrangement will be compared below to the Spiteri criteria.

4. SEED DATA

Naturally a taxonomy requires seed data – preferably a complete collection of all examples – this kind of systematic review is beyond the scope of this paper particularly as the communities of practice are still debating the definition of Web Observatories and many examples in existence may fall under a broadly accepted definition without explicitly using the term "Observatory" – making a systematic review quite challenging. As an alternative we have collated Observatory material from two sources: published papers on Observatory research and case study material from organizations expressing their own requirements for a WO. As the number of cases was deliberately small the instances were selected to cover a broad range of use type to test the theory that with the use/application of the WO features lies the key distinguishing features - hence avoiding a trivial model of Observatories about which we can say no more than they probably all store, process and output data.

We began with references to Observatories gathered from academic papers and case studies. The papers are drawn from the WWW and WebSci conferences from 2012 and 2013 and we also included insights from a small number of cases: The Trusted Data Accelerator, the SOCIAM social machine Observatory, the ODI/OpenData.Gov program and the virtual astronomical Observatory IOVA e.g.:

"The VO allows astronomers to interrogate multiple data centers in a seamless and transparent way, provides new powerful analysis and visualization tools within that system, and gives data centers a standard framework for publishing and delivering services using their data. This is made possible by standardization of data and metadata, by standardization of data exchange methods, and by the use of a registry, which lists available services and what can be done with them." (source: http://www.iova.net)

Hence a thematic/textual analysis of this material generates the raw facets - in the example above "analysis", "visualization", "registry", "publishing" etc., which can then be organized and visualised.

5. ADDING A MAP

The facets of Web Observatories that were extracted from the seed data as features/foci were listed and arranged both into groups according to the chosen method but also (noting visualisation as one of the weakness of faceted classifications) rendered as a map (see Figure 1). This figure depicts the set of features supporting all the stated requirements from the cases and thus effectively acts as a superset of the papers/cases studied. It is rendered in an implementation-neutral fashion and pre-supposes no particular hardware, storage or networking approach.

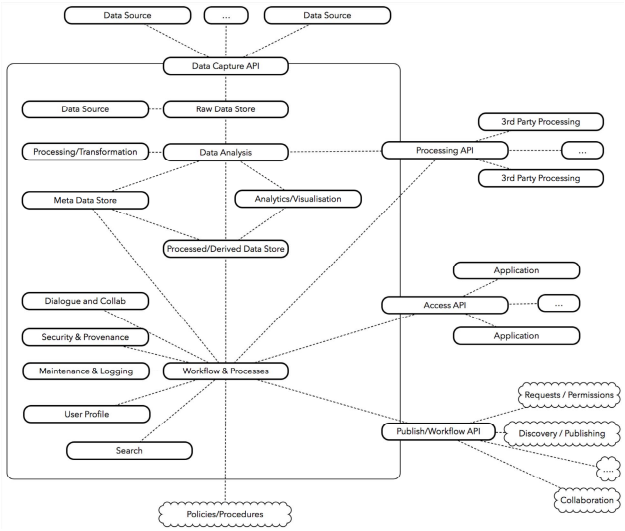

Fig 1. Concept Map of an Observatory

This representation is similar to a *concept map* (after Novak [8]) and is not intended to imply a physical design but is a representation of an Observatory in terms of which concepts that it *could* support (rather than how this would be achieved) or who would use/implement it. It can be thought of as a superset of concepts, which complements the taxonomy through the ability to visualise how certain taxonomic features may relate to one another. It should, however, be noted that not all elements of the taxonomy are currently depicted here and future work will look to extend this. In each individual concept map the border of the map represents the scope of control/authority of the Observatory owner and thus elements which cross this boundary imply "flows" requiring an interface of sorts (manual or automated). The interface in each case may correspond to a flow of data, services, communication, consensus or payment/exchange. In on-going work we are considering open source tools to capture this type of Observatory concept map notation and its flow types. This would allow sharing of machine-readable definitions including the visualization itself plus annotations, colour-coding etc. to further enhance readability/understanding. It is under investigation whether any existing tools could support both the concept map and the representation of the Taxonomy itself.

It is clear, given the stated objective to share data between Observatories, that each concept map might represent only one node in a network of Observatories engaged to address a particular question. In Brown et al. [4] we first suggested this broader nature of Observation in which each individual collection of processes may be engaged with multiple other collections in an orchestrated process. Current work is being undertaken to incorporate these processes into the model presented here.

A further benefit of the concept map is assistance during the analysis and construction of the map to eliminate any redundancy in identified facets (two features of the concept map performing the same task) and also to identify additional (missing) facets, which become may become clear through visualizing missing flows (gaps) on the map.

As Figure 1 vs. Figures 3 and 4 shows, the relationship between types of data and types of services can be more easily appreciated graphically than in two high-level facets: DATA vs. SERVICES alone.

Thus we note that data can be gathered both internally and externally, captured locally as raw data, processed/derived as synthetic data and metadata and we note also that this processing can occur locally or remotely. A set of services comprising data and/or analytics can be accessed through a service API and a set of services are foreseen around the operation of the Observatory itself including publishing the available catalogue of data and services.

We stress this is not intended to represent the "design" of an Observatory but rather to summarise the functional elements for the purposes of the taxonomy.

6. THE TAXONOMY

The source data (papers and case study documents) generated 119 foci (or features) of the Web Observatory, including features, data types, data sources etc. which were filtered and grouped into five main facets which were then further refined into 30 sub-facets. These are shown in Figures 2-4. These largely reflected the obvious groupings of the foci but use of the concept map also generated addition features/concepts and led to revised sub-groupings.

The Top-level facets are:

1. DATA
2. SERVICE
3. INTERFACES
4. PLATFORM
5. ACTORS

which we believe correspond to the mutual exclusivity and jointly exhaustive criteria required in the literature. In terms of other criteria there may be more issues around permanence of features/sources for an evolving Observatory hence we have chosen to model DATA/SERVICES at a high level so that if sub-facets such as, say, Twitter_feed [Yes/No] were to be replaced by NewService_feed [Yes/No] that this might have fewer implications than modelling a specific function or data source at the top level.

Whilst we considered that topic-focused end users of a taxonomy of Observatories are unlikely to search for an Observatory based on its technical features this may not be the case for designers or Web Science researchers looking at the "art/craft" of Observatories and hence we have retained a facet for platform features which may be extended to include platform implementation details as well as platform objectives (such as performance). This naturally led to capturing a facet around Actors/users since different users are likely to be addressing different projects with varying objectives and reaffirmed the inclusion of a collaboration (orchestration) system within an Observatory.

A searchable version of the current version of the facet list (Figure 2) has been placed on-line for the purposes of feedback/sharing at http://webscience.me/obs/web-Observatory-facets/ and we anticipate updating this visualisation as the research continues.

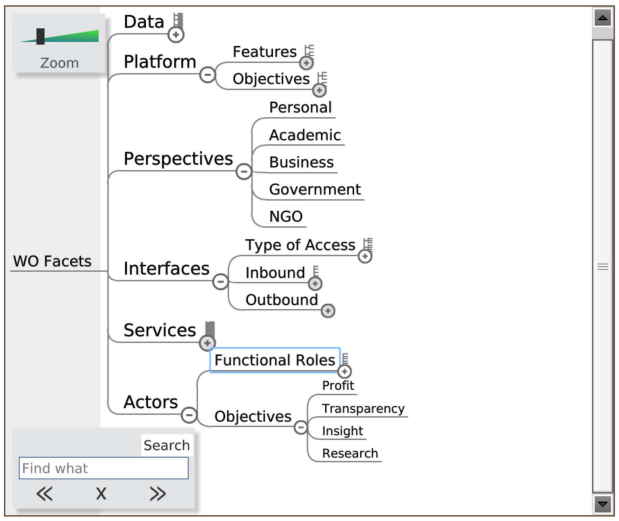

Fig 2. Online 1st, 2nd Facets for Platform, Perspectives, Interfaces and Actors

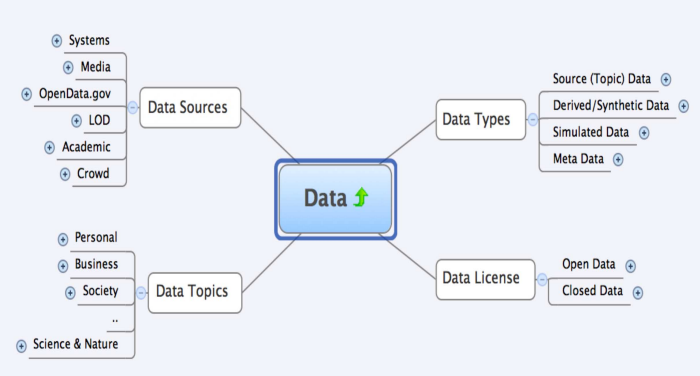

Fig 3. 1st, 2nd level Facets for Data

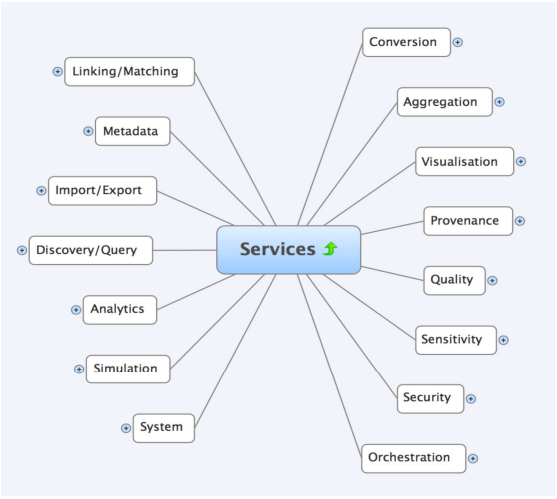

Fig 4. 1st level Facets for Services

6.1 Evaluating the work

We must engage the community of Observatory builders to determine how accurate this initial classification structure may be but of the seven criteria specified for evaluating faceted classification we would propose that we have addressed Spiteri's criteria (derived from Ranganathan and the CRG's criteria) as follows:

1. Differentiation – Top level facets are fully differentiated
2. Relevance – partially. e.g. The focus on platform details may not be relevant to all users of the classification
3. Ascertainability – partially (platform objectives such as "scalability" are poorly defined in the literature)
4. Permanence – fully – whilst sources/topics may change we feel the top-level facets will be stable.
5. Homogeneity – partially. Topic Data and metadata may be homogenous (or converted to such) within a particular classification but all OSN sources will not be functionally equivalent
6. Mutual Exclusivity – partly. Interfaces may thought to be a sub-set of Services but we have chosen to pull this out separately for the purposes of understanding WO usage.
7. Fundamental Categories – fully. None of the facets function as more general facet of the others

Finally whilst it clear that "data" may not always imply a "service", services often do imply some underlying data which they deliver or from which they are driven: examples such as Provenance and Analytical services are cases in point. In order to avoid classifying all data as a type of service we have elected to further distinguish between the types of data (shown in Figure 3) as:

1. Underlying (topic) data,
2. Derived (calculated) data,
3. Simulated data and
4. Metadata

so that we may make this distinction between the use and analysis of different data sources in the understanding that access to these data may be via services listed elsewhere in the taxonomy.

7. CHALLENGES

Issues around this work included the question of deciding between informal tagged collections and formal faceted taxonomies employing Ranganathan's stipulations around orthogonality and mutual exclusivity of facets. Whilst it may be easier to simply create tags around the Observatory definition:

<DataSources>

<Twitter>

<Facebook>

<Flickr>

</DataSources>

in effect specifying this Observatory has the following *n* data sources and making all the tags orthogonal to one another, Wilson [9] highlights problems with this when the taxonomy is implemented in software both in terms of implementation performance and semantic clarity. Instead recommends individual [yes/no] facets to distinguish between those facets, which are mutually exclusive and those, which may co-exist.

<Twitter_Source> [Yes/No]

<Facebook_Source> [Yes/No]

<Flickr_Source> [Yes/No]

Asking/Specifying on a case-by-case basis is this data source available Yes/No allows any combination of sources.

Individual facet values may be mutually exclusive (such as Data Source owner) for a particular item, but multiple Data Sources could be combined where a large number of independent (orthogonal) facets co-exist. Specifying each data source in its own facet might represent a large overhead upfront but may be required to reflect this ability to mix/combine some facets whilst restricting others. Wilson argues that tags fail to make this distinction since all tags are orthogonal to all other tags and are not exclusive.

Consider that we might want to search specifically for data over a range of topics ($T_i..T_n$) and/or across a range of sources ($S_i..S_n$) and hence the ability to efficiently implement a selection from T_i *only* in S_i versus returning all T_i across *any* S requires (argues Wilson) the facets to be separately defined. He extends the definition of the facet as:

A set of headings in which the assignment of one heading to a resource limits the assignment to that resource of other headings in the set.

Thus if we were to group all data sources under a single facet then only one could be used at a given time (think of a faceted search on the web where a single value is chosen from a drop-down). This is clearly not the intention of our WO and hence a more explicit set of facets are required (think of multiple check boxes).

Secondly, the faceted taxonomy alone does not express the relationship between facets and thus we have used an additional concept map to indicate potential process relationship between platform services and other facets (such as interfaces).

Finally we have introduced a new topic-based taxonomy as in:

Data>Data Topics>Society>OSN>

Which naturally competes with established classification schemes (such as Dewey Decimal and Library of Congress) e.g.

006.75 Specific types of multimedia systems
006.754 Online social networks

and it seems unclear at this time that a new method offers any benefits over the prior art. We leave this question to future research and offer our Data Topic Classification only as a place holder.

8. A NOTE ON SOCIAL MACHINES

In looking at Observatories we are potentially faced with a triad of perspectives: namely the data/technologies, the users and the behaviours/uses, which emerge from the interaction of the two. This lies at the heart of research into socio-technical systems and "social machines" (Shadbolt et al. at http://sociam.org). In this paper we have started with an analysis of technical themes (technologies including data) from systems that have already been planned/designed to meet some stated user need. We could equally have started from the perspective of the Actors/Users to analyse how technical Observatory solutions may be socially constructed. There are well-established schools in this area such as Actor-Network Theory and whilst this kind of analysis is beyond the scope of the current paper we have deliberately included the concept of Actors/Users in the Taxonomy in recognition of the importance of this fact. Whilst not all social machines are Observatories it could be argued that all Observatories are social machines and, as such, any taxonomy must provide for the description of the social element of that social machine. Given the complexities of orchestrating multiple services over multiple data sources across multiple Observatories it is far from given that all operations and services within an Observatory must be purely technological/automated in nature rather than manual/social computations. We note analogues in the financial services sector from such providers as Thomson Reuters and Bloomberg where diverse data sources are combined/mapped to provide a range of service products. In many cases there remain manual interventions in the management and operation of these services.

9. CONCLUSIONS

We have shown that Observatories may be classified via a flexible faceted approach allowing for extensibility not only within the definition of what Observatories are but also in terms of social perspective of what they are for. We propose that a content/innovation perspective (addressing what people are trying to achieve) is likely to promote better understanding of how WO's are developed and extended rather than starting from a purely physical/technical perspective which describes how they do it.

Naturally the efficiency and scalability of Observatories relies heavily on sound technical/architectural choices for storage, querying and analytics but in terms of a *functional* definition we believe that, like consumers of electricity, Observatory users may be less concerned with how their power is generated than with the fact that it is reliable (trustworthy), available and compatible with the devices they want to use.

This work is parallel to Shadbolt et al. [10] who are attempting to classify social machines (of which the Observatory is an example). Observatories are proposed as a tool to study (other) social machines.

Future work in this area comprises the further development of the simple concept map to address process interactions and the Actor/User dimension of Observatories. We plan to extend the seed data to a wider systematic review of Web-based repositories that might be considered Observatories and finally consider the technical implementations of the concept map and taxonomy in a tool supporting a suitable mark-up schema such as XFML (eXchangeable Faceted Metadata Language). This last step could enable prototype search/discovery methods to dovetail with other research on Observatory discovery. It should also be noted that the current model may requires extensions to consider larger scale interactions between multiple WO's.

10. ACKNOWLEDGEMENTS

This work is supported under SOCIAM: The Theory and Practice of Social Machines. The SOCIAM Project is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J017728/1 and comprises the Universities of Southampton, Oxford and Edinburgh. Also by the Web Science Trust and the EPSRC grant number EP/K503150/1.

11. REFERENCES

[1] B. Kwasnick, "The Role of Classification in Knowledge Representation and Discovery." [Online]. Available: https://www.ideals.illinois.edu/bitstream/handle/2142/8263/librarytrendsv48i1d_opt.pdf. [Accessed: 10-Jan-2014].

[2] L. Spiteri, "A Simplified Model for Facet Analysis," Canadian Journal of Information and Library Science v23, 1-30 (April-July 1998). [Online]. Available: http://iainstitute.org/en/learn/research/a_simplified_model_for_facet_analysis.php [Accessed: 10-Jan-2014].

[3] R. W. Morshead, "Taxonomy of Educational Objectives Handbook II: Affective Domain," *Studies in Philosophy and Education*, vol. 4, no. 1, pp. 164–170, 1965.

[4] I. Brown, W. Hall, and L. Harris, "From search to observation," *International World Wide Web Conferences Steering Committee*, pp. 1317–1320, May 2013.

[5] S. R. Ranganathan, "Prolegomena to Library Classification," 3rd Edition, Asia Publishing House, 1967.

[6] B. C. Vickery and A. Y. Oficinas, "Faceted classification: a guide to construction and use of special schemes," Aslib, 1960.

[7] W. Denton, "How to Make a Faceted Classification and Put It On the Web," pp. 1–12, Jan. 2003. [Online]. Available: http://www.miskatonic.org/library/facet-web-howto.html. [Accessed: 10-Jan-2014].

[8] Novak, J. D. & A. J. Cañas, "The Theory Underlying Concept Maps and How to Construct and Use Them", Technical Report IHMC CmapTools 2006-01 Rev 01-2008,

Florida Institute for Human and Machine Cognition, 2008, available at: http://cmap.ihmc.us/Publications/ResearchPapers/TheoryUnderlyingConceptMaps.pdf. [Accessed: 12-Jan-2014]

[9] T. Wilson, "The strict faceted classification model." [Online]. Available: http://facetmap.com/pub/strict_faceted_classification.pdf. [Accessed: 12-Jan-2014].

[10] N. R. Shadbolt, D. A. Smith, E. Simperl, M. Van Kleek, Y. Yang, and W. Hall, "Towards a Classification Framework for Social Machines," pp. 1–7, SOCM2013: Workshop on Theory and Practice of Social Machines, WWW2013 2013, Rio de Janeiro, Brazil, April 2012