The Role of Community-Driven Data Curation for Enterprises

Title	The Role of Community-Driven Data Curation for Enterprises
Author(s)	Curry, Edward;Freitas, Andre;O'Ríain, Seán
Publication Date	2010
Publisher	Springer US

The Role of Community-Driven Data Curation for Enterprises

Edward Curry, Andre Freitas, and Sean O'Riáin

Abstract With increased utilization of data within their operational and strategic processes, enterprises need to ensure data quality and accuracy. Data curation is a process that can ensure the quality of data and its fitness for use. Traditional approaches to curation are struggling with increased data volumes, and near real-time demands for curated data. In response, curation teams have turned to community crowd-sourcing and semi-automated metadata tools for assistance. This chapter provides an overview of data curation, discusses the business motivations for curating data and investigates the role of community-based data curation, focusing on internal communities and pre-competitive data collaborations. The chapter is supported by case studies from Wikipedia, The New York Times, Thomson Reuters, Protein Data Bank and ChemSpider upon which best practices for both social and technical aspects of community-driven data curation are described.

1 Introduction

Using data and quantitative analysis to support decision making is a growing trend within the business environment with many companies reaping significant benefits [1]. However, one of the major pitfalls in data driven decision making is poor quality data. In its December '09 issue The Economist highlighted the problems

Edward Curry
Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland, e-mail: ed.curry@deri.org

Andre Freitas
Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland, e-mail: andre.freitas@deri.org

Sean O'Riáin
Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland, e-mail: sean.oriain@deri.org

D. Wood (ed.), *Linking Enterprise Data*, DOI 10.1007/978-1-4419-7665-9_2,

some banks have with data quality. "The same types of asset are often defined differently in different programs. Numbers do not always add up. Managers from different departments do not trust each other's figures". One bank official noted "some figures were not worth the pixels they were made of", highlighting inaccurate figures make it very difficult to manage operational concerns such as investments exposure and risk.

Making decision based on incomplete, inaccurate, or wrong information can have disastrous consequences. Decision making knowledge workers need to have access to the right information and need to have confidence in that information. Data curation can be a vital tool to ensure knowledge workers have access to accurate, high-quality, and trusted information that can be reliably tracked to the original source, in order to ensure its credibility.

This chapter discusses the business motivations for curating data and investigates the role of community-based data curation, especially pre-competitive collaborations. The chapter provides an overview of data curation supported by case studies from Wikipedia, The New York Times, Thomson Reuters, PDB and ChemSpider along with best practices for both social and technical aspects of data curation.

2 The Business Need for Curated Data

The increased utilization of data, with a wide range of key business activities (including business intelligence, customer relationship management and supply chain management) has created a data intensive landscape and caused a drastic increase in the sophistication of enterprise data infrastructures. One of the key principles of data analytics is that the quality of the analysis is dependent on the quality of the information analyzed. However, within operational enterprise systems, there is a large variance in the quality of information. Gartner recently estimated that more than 25 percent of critical data in the world's top companies is flawed[1]. Uncertainty over the validity of data or ambiguity in its interpretation can have a significant effect on business operations, especially when it comes to the decision making process. When making decisions or interpreting business intelligence reports or dashboards, business executives must be able to assess the quality of the data they are using.

Perception of data quality is highly dependent on the fitness for use [2]; being relative to the specific task that a user has at hand. Data quality is usually described by a series of quality dimensions which represent a set of desirable characteristics for an information resource (see [3] for a survey of the main data quality frameworks). The following data quality dimensions are highly relevant within the context of enterprise data and business users:

- **Discoverability & Accessibility:** Addresses if users can find the data and then access it in a simple manner. It is important to facilitate users in their search for

[1] Gartner, Inc press release. "'Dirty Data' is a Business Problem, Not an IT Problem, says Gartner," March 2, 2007

information. Data curation can streamline the users search for data by storing and classifying it in an appropriate and consistent manner.

- **Completeness:** Is all the requisite information available? Does it need to be cleansed for errors and duplications? Are there any omissions (values and whole records) from the data? In some cases, missing data is irrelevant, but when the information that is missing is critical to a specific business process, completeness becomes an issue. Data curation can be used to conduct data audits and develop data-retention criteria to improve the completeness of and remove duplications from enterprise data. Curation can also be used to provide the wider context of data by linking/connecting related datasets.
- **Interpretation:** Is the meaning of data ambiguous? Reaching a common interpretation of information is a challenging task. Humans can have different underlying assumption that can significantly affect the way they interpret data. While no technology is capable of eliminating misinterpretations, data curation can ensure that any assumptions made during information collection (and calculations) are made explicit, to minimize misunderstandings.
- **Accuracy:** Is the data correct? Unreliable data has drastically reduced usefulness for the enterprise. Data curation can be used to ensure that data correctly represent the "real-world" values it models.
- **Consistency:** Does the data contradict itself? Are values uniform across datasets? Inconsistent data can introduce significant difficulty for organizations attempting to reconcile between different systems and applications. Data curation can be used to ensure that data is created and maintained to ensure it uses standardized definitions, calculations, terms, and identifiers in a consistent manner.
- **Provenance & Reputation:** Is the data legitimate? Where did the data come from? Can it be reliably tracked back to the original source? Is the source of the data highly regarded? Data provenance can be used to assess the trustworthiness and quality behind data production and delivery. When judging the quality of data the reputation of the data sources and/or data producers (organizations and individuals) can be a critical factor. Data curation activities could be used to both track the source of the data and determine their reputation. Reputation can also include the objectivity of the source/producer. Is the information unbiased, unprejudiced, and impartial? Or does it come from a reputable but partisan source?
- **Timeliness:** Is the information up-to-date? Data curation can be used to determine how up-to-date data is, with respect to the task at hand.

By improving these quality dimensions, data curation can increase the credibility and trust of data that passes through the curation process. However, not all enterprise data should be curated. Data curation is suited to knowledge-centric data rather than transactional operations data. It is important to note that not all knowledge-centric enterprise data should be curated, given the associated effort and cost required it would not make sense to do so. Each organization must identify what data would benefit most from curation, and determine if potential returns would support the required investment.

Data governance is an emerging field that is a convergence of data quality, data management, business process management, and risk management which surrounds the handling of data in an organization. A full discussion of data governance is outside the scope of this chapter, but it is important to know that data curation is a complimentary activity that can form part of an overall data governance strategy for the organization.

3 Data Curation

Digital curation is the process of establishing and maintaining a trusted body of digital information within long term repositories for current and future use by researchers, scientists, historians, and scholars generally. Specifically digital curation is defined as the selection, preservation, maintenance, collection, and archiving of digital assets [4].

Data curation, a subset of digital curation, is the active management and appraisal of data over its life-cycle of interest. In the same manner in which a museum curator ensures the authenticity and organization of a museum's collection, data curation is performed by expert curators responsible for improving the accessibility and quality of data. Data curators (also known as bio curators, scientific curators, or data annotators) are recognized as the "museum cataloguers of the Internet age" [5] who have the responsibility to ensure that data is trustworthy, discoverable, accessible, reusable and fit for its current intended use.

3.1 How to Curate Data

The Digital Curation Centre[2] provides extensive support services on digital curation and preservation to the established 'traditional' curation community outlining initiatives towards curation standards, procedures and tools. For the purposes of this chapter, we will provide a high-level overview of some key questions must be addressed to setup a curation process within an organization.

An obvious starting point is to identify the business use case for creating a curated dataset. Typically a curation effort will have a number of motivations to curate data, including improving accessibility, quality, or tracking provenance.Once clearly established, one can start to define "how" the data will be curated. There is no single correct way to curate data and there are many ways to setup a data curation effort. The major factors that will influence the curation approach include the:

- Quantity of data to be curated (include new and legacy data),
- Rate of change of the data,
- Amount of effort required to curate the data, and the

[2] Digital Curation Centre, http://www.dcc.ac.uk/ Last Accessed on 8th June 2010

- Availability of experts.

Once these factors are determined, an estimation of the work required to curate the data can be made. If dealing with an infrequently changing small quantity of data (<1,000 records) with minimal curation effort (minutes per record), curation could be easily undertaken by an individual. However, once the number of records enters the thousands, a curation group/department with a formal process has to be considered. Curation departments can deal with large curation efforts, especially when they utilize software to support the curation process, but there is a limit to their scalability. When curating large quantities of dynamic data (>million records) even the most sophisticated curation department can struggle with the workload. An approach to curate data on this scale is to utilize crowd-sourcing *community-based curation* in conjunction with computer-based curation. One popular crowd-sourcing technique is *sheer curation* that integrates curation activities within the normal workflow of those creating and managing the target data. These curation approaches are not mutually exclusive and can be used in conjunction to curate data. These blended approaches are proving to be successful.

3.1.1 Setting up a Curation Process

- **Identify what data you need to curate:** Will you be curating newly created data and/or legacy data? How is new data created? Do users create the data, or is it imported from an external source? How frequently is new data created/updated? What quantity of data is created? How much legacy data exists? Where does the legacy data reside? Is it stored within a single source, or scattered across multiple sources.
- **Identify who will curate the data:** Curation activities can be carried out by individuals, departments or groups, institutions, communities, etc.
- **Define the curation workflow:** How will curation activities be carried out? The curation process will be heavily influenced by the previous two questions. The two main methods to curate data are a curation group/department or a sheer curation workflow that enlists the support of users.
- **Identity the most appropriate data-in and data-out formats:** What is the best format for the data to be expressed? Is there an agreed standard format within the industry or community? Choosing the right data format for receiving data and publishing curated data is critical; often a curation effort will need to support multiple formats to ensure maximum participation.
- **Identify the artifacts, tools, and processes needed to support the curation process:** A number of artifacts, tools, and processes can support data curation efforts, including workflow support, web-based community collaboration platforms. A number of algorithms exist to automate or semi-automate curation activities [6] such as data cleansing[3], record duplication and classification algorithms [7] that can be used within sheer curation.

[3] Detecting and correcting (or removing) inaccurate or corrupt records

4 Community-based Curated Enterprise Data

Data curation can be a time consuming and difficult task. Often the effort required to curate anything but a trivial dataset is beyond the capability and capacity of a single individual. As such, it is often within the interest of individuals to join community efforts to curate the data. By becoming part of a community, participants are able to share the costs, risks and technical challenges, while benefiting from the wisdom of the community and the network effect for their curated dataset. One of the most popular community curated datasets is the Wikipedia online encyclopedia, which will be analyzed as one of the case studies in this chapter.

Many enterprises use community-based approaches to meet their data curation needs, which have proved very successful for knowledge centric data curation. Depending on the requirements of the data, an enterprise can utilize an internal community or participate with an external community to collaboratively curate data.

In order to determine the right model, one must consider a number of issues, such as:

- What the purpose of the community is?
- Who participates within the community?
- Will access to the curated dataset be publicly available? Or restricted?
- Is the curation process open to public participation? Or limited to a selected curation group?
- What is the community governance model?

Once these questions are answered one can start to determine what community model will best suit. Two popular models are *internal corporate communities* and *external pre-competitive communities*.

4.1 Internal Corporate Community

Enterprises have started to tap the potential of their workforce to assist within the data curation process. Internal corporate communities can be utilized to curate competitive enterprise data that will remain internal to the company, although this may not always be the case (e.g. product technical support and marketing data). Internal communities often work in conjunction with a formal curation department and their governance will typically follow the organization's internal governance model.

A typical approach is to create a department that consists of curation experts that can work in conjunction with subject matter experts to curate data, after it has been created, in a post hoc manner. This "traditional" form of data curation has proved to be very successful. However, the post-hoc nature of the approach creates a delay in the availability of curated data. With business relying more on data in their day-to-day operations, there is a need to reduce the time taken to make curated data available. Making the situation even more challenging is the increased quantities of data that need to be curated. Data curation teams have found it difficult to scale

the traditional approach and have turned to crowd-sourcing [8] and automated/semi-automated metadata annotation tools to assist the curation process [6].

With the increased use of online collaboration tools and the need to curate larger amounts of data, many enterprises have employed decentralized approaches to data curation by turning to internal communities of users to curate data. Often the curation task can be done as the data is created. *Sheer curation*, or *curation at source*, is an approach to data curation where lightweight curation activities are integrated into the normal workflow of those creating and managing data and other digital assets. The results of the sheer curation process can be made immediately available. Sheer curation activities can be as simple as vetting or "rating" the results of a categorization process performed by a curation algorithm. Sheer curation activities can also be combined with the activities of a post hoc curation department to provide more sophisticated curation activities. These blended approaches to data curation allow more immediate access to curated data while also ensuring the quality control that is only possible with an expert curation team.

4.2 External Pre-competitive Communities

For data that must remain private to an enterprise, for competitive reasons, internal communities are best suited. However, a growing trend is for enterprises to participate within external data curation communities where the data is deemed to be pre-competitive [8, 9]. Many organizations, both commercial and non-profit, have come together to build sustainable data curation communities that share costs, risks, and technical challenges between members. These communities leverage a larger user base for crowd-sourcing and provide a distinct advantage to improve the wisdom of the crowds [8].

Pre-competitive collaboration is a well-established technique, with a number of industries realizing the benefits of an open innovation model for collaboration. Notable examples are the Airbus consortium of European aircraft manufacturers, the Sematech consortium of US semiconductor manufacturers, and banks working together to launch Visa and MasterCard.

A typical company will leverage its propriety data for competitive advantage. However, many companies also utilize common data that does not provide any competitive advantage. While this data has little potential for differentiation, the company must still invest in maintaining and curating the data. Often many companies will duplicate this effort in-house, incurring the full-cost of maintaining the dataset. In order to avoid this unnecessary cost, companies can collaborate within pre-competitive initiatives.

Pre-competitive data is information that can be shared without conferring a commercial advantage to a competitor. Pre-competitive curation collaboration activities between organizations can help to overcome decreasing budgets by reducing the costs required to provide and maintain data, while increasing the quantity, quality and access to non-competitive data. Company participation within a pre-competitive

community can take the form of a direct monetary contribution, personnel contribution, and/or by donating datasets. The common data curation tasks can be carried out once within the public domain rather than multiple times within the private domains of each company.

Participation within a pre-competitive curation community allows participating companies to focus on value-adding competitive activities such as data analysis and data exploration, the Protein Data Bank and ChemSpider case studies being examples of this. Another example is the Pistoia Alliance, a consortium in the Pharmaceutical industry [4]. The objective of these communities is to move the "competitive onus" from novel data to novel algorithms by shifting the "emphasis from 'proprietary data' to a 'proprietary understanding of data'"[10]. The scope of the community can also extend beyond just data curation tasks to include collaboration for developing common pre-competitive software infrastructures for manipulating and storing the data.

Two popular community models are: *organization consortium* and *open community*. An *organization consortium* is a type of community which operates like a private democratic club where participating organizations collaborate on curation activities. The benefit is being able to share risks, costs and technical challenges while also sharing any potential IP created. Consortiums are usually a closed community where members are invited based on their skill set to provide a contribution to curation activities. The availability of the resulting output data may be publicly available or limited to the consortium members. Consortiums follow a democratic process; however the voting rights of each member may reflect the level of investment they make within the consortium. Within these scenarios larger players may become the leaders of the consortium.

Within an *open community* everyone can participate. The founder(s) of the community defines the desired curation activity and seeks public support from a potential unlimited number of participants who feel they have the skills to provide or contribute to curation activates. Wikipedia, Linux, and Apache are good examples of large open communities where anyone can contribute.

5 Case Study: Wikipedia - The World Largest Open Digital Curation Community

Wikipedia is an open-source encyclopedia, built collaboratively by a large community of web editors. The success of Wikipedia as one of the most important sources of information available today still challenges existing models of content creation. As of March 2010, Wikipedia counted more than 19,000,000 articles, with over 3,200,000 in the English language. Wikipedia covers near 270 languages and counts with more than 157,000 active contributors. Previous investigations showed the evidence that both the accuracy [11] and stylistic formality [12] are equivalent to

[4] The Pistoia Alliance http://www.pistoiaalliance.org/ Last Accessed on 8th June 2010

resources developed in expert-based closed communities such as the Columbia and Britannica encyclopedias. Despite the fact that the term 'curation' is not commonly addressed by Wikipedia's contributors, the task of digital curation is the central activity of Wikipedia editors, who have the responsibility for information quality standards.

Wikipedia uses a wiki as its main system for content construction. Wikis were first proposed by Ward Cunningham in 1995 and allow users to edit contents and collaborate on the Web more efficiently. MediaWiki, the wiki platform behind Wikipedia, is already widely used as a collaborative environment inside organizations[5]. Important cases include Intellipedia, a deployment of the MediaWiki platform covering 16 U.S. Intelligence agencies[6], and Wiki Proteins, a collaborative environment for knowledge discovery and annotation [13].

The investigation of the collaboration dynamics behind Wikipedia can highlight important features and good practices which can be applied to different organizations. Our analysis focuses on the curation perspective and covers two important dimensions: *social organization* and *artifacts, tools & processes* for cooperative work coordination. These are key enablers that support the creation of high quality information products in Wikipedia's decentralized environment.

5.1 Social Organization

One important feature behind the Wikipedia initiative is the idea of lowering barriers for new contributors, by allowing any user, without prior registration, to edit its contents. What one would have expected to lead to a chaotic scenario, proved to be, in practice, a highly scalable approach for high quality content creation on the Web. Wikipedia relies on a simple but highly effective way to coordinate its curation process and accounts and roles are in the base of this system.

Wikipedia has four main types of accounts: (a) anonymous users - which are identified by their associated IP address, (b) registered users - users with an account in the Wikipedia website, (c) administrators - registered users with additional permissions in the system and (d) bots - programs that perform repetitive tasks. All users are allowed to edit Wikipedia contents. Administrators, however, have additional permissions in the system.

For the definition of the central roles in the curation process we refer to a subset of the roles identified by Stivilia [14]: *editor, administrator, bureaucrat, steward, arbitrator, mediator* and *bots*. The roles can provide a more clear perspective of the tasks that accounts perform, where the arbitrator and the mediator roles can be performed by bureaucrats and administrators. Bureaucrats and stewards are special

[5] MediaWiki Testimonials, http://www.mediawiki.org/wiki/Sites_using_MediaWiki/corporate Last Accessed on 8th June 2010

[6] CIA.gov Featured Article, "Intellipedia Celebrates Third Anniversary with a Successful Challenge", https://www.cia.gov/news-information/featured-story-archive/intellipedia-celebrates-third-anniversary.html Last Accessed on 8th June 2010

types of administrators who can perform additional functions while acting on consensus. For a more detailed description of the accounts and roles types the reader is directed to Stivilia [14].

Wikipedia can provide important insights about the social dynamics of open collaboration on the Web. Kollock concluded that the incentives behind open collaboration are based on the expectation of future reciprocity from the community, the improvement of one's reputation and the sense of efficacy (contributing effectively to a meaningful project) [15]. While Bryant et al. [16] investigated the transformation of roles in the process of becoming a Wikipedian (a Wikipedia editor). By interviewing Wikipedians, the authors observed that, over time, the focus of editors commonly changed from curators of a few articles in topics where they were familiarized to a more global curation perspective, motivating and enforcing the quality assessment of Wikipedia contents as a whole.

5.2 Artifacts, Tools and Processes

Wikipedia makes use of different artifacts, tools and processes to provide editors guidance in the digital curation process. In contrast to other environments where information quality is enforced by the application of restrictive permission mechanisms, Wikipedia provides a minimal and effective infrastructure as described below.

- **Wiki Article Editor (Tool):** A wiki is a website which allows users to easily create, edit and publish contents in web pages through the use of a WYSIWYG or markup text editor.
- **Talk Pages (Tool):** Talk pages represent a public arena for discussions around Wikipedia resources. Talk pages are used with the purpose of discussion lists where each editable resource has an associated Talk page. The work of Viégas et al. [17] provide a detailed analysis of the talk pages role in the coordination of the edition of resource contents. Talk pages serve as a multichannel tool allowing users to request/suggest editing coordination, request for information, reference vandalism, reference Wikipedia guidelines and policies, reference internal Wikipedia resources, write off-topic remarks, make polls, request peer review, define status through information boxes, post images, together with other minor uses.
- **Watchlists (Tool):** Every user can put a Wikipedia resource in a watchlist in order to receive notifications of changes of the state of a specific resource. Watchlists help curators to monitor actively the integrity and quality of the set of resources which they contribute.
- **Permission Mechanisms (Tool):** Users with administrator status have the permission to perform critical actions inside the system such as remove pages, grant administrative permissions for new users.
- **Automated Edition (Tool):** Bots are automated or semi-automated tools that perform repetitive tasks over the Wikipedia contents.

- **Page History and Restore (Tool):** The historical trail of changes of a Wikipedia Resource can be accessed in the page history. Editors with certain administrator status can restore the previous status of a resource.
- **Guidelines, Policies & Templates (Artifact):** Resources including 'The Perfect Article'[7], Featured Articles[8] and Layout[9], define the curation guidelines for editors to assess the quality of an article. In addition, a comprehensive set of policies[10], covering every critical aspect of the editorial process behind Wikipedia, are defined.
- **Dispute Resolution (Process):** Disputes between editors over the contents of an article can lead to different dispute resolution mechanisms.
- **Article Edition, Deletion, Merging, Redirection, Transwiking, Archival (Process):** These processes describe the curation actions over Wikipedia resources.

5.3 DBPedia - Community Curated Linked Open Data

Wikipedia provides document-centric access access to information. DBPedia, on the other hand, provides direct access to data through its comprehensive infrastructure of concept URIs, their definitions and basic types.

As of March 2010, DBPedia counted 3.4 million entities and 1 billion RDF triples. DBPedia inherits the massive volume of curated data available at Wikipedia and indirectly uses its wiki infrastructure as a curation platform. Since DBpedia has a broad scope of entities covering different areas of the human knowledge, it is natural hub for connecting datasets, where external datasets could link to its concepts.

The DBPedia knowledge base is built using the information present in well defined links inside the article and infobox-specific properties. The general properties include a label, an abstract, a link to the Wikipedia Article, links to related DBPedia entities, links to external Web resources, a link to an image of the concept and geo-coordinates. Infobox specific properties are mapped using two types of extractors: generic infobox extraction, which build predicates directly from the pairs of attribute-value present on infoboxes and mapping-based infobox extraction, which uses a manually created ontology (170 classes, 720 properties) built from the 350 most frequent infoboxes templates to map the attribute-value pair to the ontology terms. The reader is referred to [18] for a more detailed discussion about the extraction process behind DBPedia.

The use of a wiki as a collaborative platform for the creation and maintenance of lightweight ontologies is covered by Hepp et al. in [19], which also found that Wikipedia can provide high quality and stable concept identifiers. Hepp also high-

[7] http://en.wikipedia.org/wiki/Wikipedia:The_perfect_article Last accessed on 8th June 2010

[8] http://en.wikipedia.org/wiki/Wikipedia:Featured_articles Last accessed on 8th June 2010

[9] http://en.wikipedia.org/wiki/Wikipedia:Layout Last accessed on 8th June 2010

[10] http://en.wikipedia.org/wiki/Wikipedia:List_of_policies Last accessed on 8th June 2010

lights that ontologies work not only as a formal representation of a specific domain but also as a community contract over this representation. By having a larger community of contributors, the collaboratively created ontology is more likely to express this social agreement. One additional positive point is the fact that, by being a widespread and popular technology, wikis can lower the entry barriers for collaborative data curation. In this context, despite being targeted to document curation, wikis can also support data curation.

6 Case Study: The New York Times - 100 Years of Expert Data Curation

The New York Times (NYT) is the largest metropolitan and the third largest newspaper in the United States. The Times website, *nytimes.com*, is ranked as the most popular newspaper website in the United States and is an important source of advertisement revenue for the company. The NYT has a rich history for curation of its articles and its 100 year old curated repository has ultimately defined its participation as one of the first players in the emerging Web of Data.

6.1 Data Curation

The history of data curation in The New York Times dates back to 1913 when, fearing competition with the New York Sun, the publisher and owner Adolph S. Ochs decided to provide a set of additions to the newspaper. One of these additions was the New York Times Index, an ambitiously organized catalog of articles titles and summaries (containing the issue, date and column of the original article), published in a period of time, categorized by subject and names. The Index was introduced on a quarterly basis at first, being later produced on an annual basis. With the introduction of the NYT Index, the typically transitory content of the newspaper became an important source of searchable historical data, often used to settle historical debates.

In order to create a high quality catalog, an Index Department was created, marking the start of a systematic data curation and cataloguing effort over the NYT resources (since 1851 the NYT already had a low quality index for internal use). In the following years, the Index Department developed a comprehensive catalog using a controlled vocabulary covering subjects, personal names, organizations, geographic locations and titles of creative works (books, movies, etc), linked to articles and their summaries. As of March 2010 the Index Department is 15 people strong.

The process of consistently and accurately classifying news articles over a large period of time can pose numerous challenges. Over the time span of 100 years the keywords used to express subjects may show some variance due to cultural or legal constraints. The identities of some entities, such as organizations and places, changed over time. Other entities, in particular names derived from non-Latin al-

phabets may show different lexical expressions. In addition, the NYT controlled vocabulary grew to hundreds of thousands of categories adding considerable complexity to the classification process.

With the increase of importance of the Web to the NYT, there was a need to improve the categorization of online contents. The curation carried out by the Index Department in library-time (days to weeks) was not suitable for real-time demands of online publishing. The print versions of the paper could handle a next-day index, but nytimes.com needed a same-day index. To meet this challenge, the NYT introduced a two stage curation process where the editorial staff performed best-effort semi-automated sheer curation at the point of online publication with the Index Department following up with long-term accurate classification and archiving. This blended approach provides the best of both worlds, with the non-expert journalist's curators providing instant accessibility to online users, and the Index Department providing long-term high-quality expert curation in a "trust but verify" approach.

The editorial staff of the New York Times consists of several hundreds journalists who work as first level curators in the content classification process. Two *taxonomy managers* review the work of the first level curators, providing constant feedback into the classification process.

The basic workflow (see Figure 1) of the first level curation starts with an article getting out of the newsroom. Using a Web application, a member of the editorial staff submits the new article through a rule based information extraction system (in this case, SAS Teragram[11]). Teragram uses a set of linguistic extraction rules which are created by the taxonomy managers based on a subset of the controlled vocabulary used by the Index Department. Teragram suggests tags based on the Index vocabulary that can potentially describe the content of the article. The member of the editorial staff then selects the terms that better describe the contents and inserts new tags if necessary. The classification is reviewed by the taxonomy managers and the content is published online. In a later stage the article receives a second level curation by the Index Department, which appends additional tags and a summary of the article to the stored resource.

6.2 Publishing Curated Linked Data

In 2009 the NYT announced the publication of a subset of nearly 10,000 tags of its indexing vocabulary as Linked Open Data (LOD), becoming one of the early companies to open their data as Linked Data. The NYT LOD initiative[12] intends to expand its vocabulary coverage to a larger set of the index.

As of March 2010, the NYT dataset consists of people, organizations and locations. The published data is complemented by the NYT Restful API, where appli-

[11] SAS teragram http://www.teragram.com Last Accessed on 8th June 2010

[12] The NYT Linked Open Data http://data.nytimes.com, Accessed on 9th March 2010

cation developers can use a set of different search services to consume data about articles, movies, best sellers, Congress votes, real estate, among other uses.

The NYT LOD initiative inherits the quality that is the consequence of almost 100 years of investment in careful data curation. The publication of its curated Linked Data unleashes a set of potential benefits to the NYT including: improvement of the online traffic by third party data usage, lowering of the cost of development of new applications for different verticals inside the website (e.g. movies, travel, sports, books), creation of better online contents by links and the potential for Search Engine Optimization.

7 Case Study: Thomson Reuters - Data Curation, a Core Business Competency

Thomson Reuters is an information provider company created by the acquisition of Reuters by Thomson Corporation in 2008. In 2009 the company had over 50,000 employees and a commercial presence in more than 100 countries. Thomson Reuters business is focused on the provision of specialist curated critical information and information-based services which can enable strategic decision making in different domains, including Healthcare, Science, Financial, Legal and Media. Thomson Reuters is among the early corporate adopters of Semantic Web technologies, progressively incorporating these on its data and services.

Thomson Reuters utilize Semantic Web Technologies to provide better contextualized, meaningful, interoperable and machine readable contents to its customers. Since the main customers of Thomson Reuters are information consumers, Semantic Web Technologies are seen as a way to bring information with integrated context to end users. The acquisition of ClearForest (by Reuters in 2007), a company focused on information extraction through Natural Language Processing, shows solid evi-

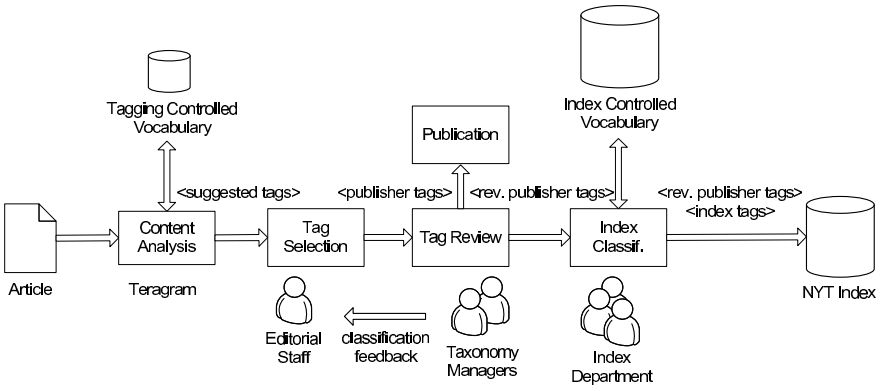

Fig. 1 The NYT article classification curation workflow

dence that improving the organization and semantics of the information can provide a competitive edge for the company

7.1 Data Curation

The objective of data curation at Thomson Reuters is to select the most relevant information for its customers, classifying, enriching and distributing it in a way that can be readily consumed. The curation process at Thomson Reuters employs thousands of curators working over approximately 1000 data sources. In the curation process, automatic tools provide a first level triage and classification which is further refined by the intervention of human curators. A curator inside Thomson Reuters is a specialist in a specific domain, who collects, aggregates, classifies, normalizes and analyzes the raw information coming from different data sources.

Since the nature of the information analyzed at Thomson Reuters is typically high volume and near real-time, data curation is a big challenge inside the company and the use of automated tools plays an important role in this process. One of these tools, OneCalais, is a platform which uses Natural Language Processing (NLP) over unstructured text to automatically derive tags for the analyzed content, enriching it with machine readable structured data. The tags enrich the original text with the general category of the analyzed contents, while also providing a description of specific entities (places, people, events, facts) which are present in the text. OneCalais was a product developed at ClearForest.

Since 2008 OpenCalais, a free public version of the extraction service provided by OneCalais, was made available to the public. By March of 2010 OpenCalais had over 20.000 users executing an average of over 4 million transactions per day. The corporate users of OpenCalais includes CNET, CBS Interactive, The Huffington Post, The Powerhouse Museum of Science and Design, which use the system as a platform to reduce the workload involved in the classification of digital collections.

Both OneCalais and OpenCalais use the Linked Data principles[13] to describe entities. Every entity inside the systems has a de-referenceable URI. From the perspective of Thomson Reuters business, Linked Data can provide a high impact data distribution strategy. However, deploying Linked Data to corporate customers brings additional challenges. Business users will not rely on an unmanaged linked data ecosystem. Thomson Reuters address part of this problem creating its own 'branded' Linked Data containing information about companies, geography, etc.

[13] T. Berners-Lee, "Linked Data Design Issues.", http://www.w3.org/DesignIssues/LinkedData.html Last Accessed on 8th June 2010

8 Case Study: ChemSpider - Open Data Curation in the Global Chemistry Community

ChemSpider[14] is a search engine that provides free service access to the structure centric chemical community. Available since 2007 ChemSpider has collated over 300 data sources from chemical vendors, government databases, private laboratories and individuals, providing access to 25 million records. Used by chemists for identifier conversion and properties predictions, its datasets are also heavily leveraged by chemical vendors and pharmaceutical companies as pre-competitive resources for experimental and clinical trial investigation. Pharmaceutical companies are starting to realize the benefits of the open data model and contribute in kind, Glaxo Smith Kline[15] is an example of an enterprise intent on making its proprietary malaria dataset of 13,500 compounds available for community consumption.

Using the Open Community model, ChemSpider distributes its curation activity across its community using crowd-sourcing[16] to accommodate massive growth rates and quality issues. Integrating online services (e.g. PubMed[17], Google Scholar, Google Books and Microsoft's academic search) allowed ChemSpider move towards an environment that provides all additional required resources such as patent structured search. In addition, The Concept Web Alliance, a partner of the ChemSpider Initiative, is looking to linked data as a strategy to organize scientific data on the Web.

Driving its community vision was the provision of an environment where the community could participate in data curation, and validation that would help the chemical structure community to solve problems. Anthony Williams, ChemSpiders' VP of Strategic Development, attributes successful community participation levels to engagement and feedback through the use of social networking (e.g. blogs, forums, twitter, friend feed). The interactions, Anthony noted, led to better understanding of the community needs and an accommodation shift on ChemSpider's part which helped guide the project path. In curation terms, gaining access to knowledge skills and understanding that otherwise would not have been possible, proved to be critical.

---

[14] http://www.chemspider.com/ Last Accessed on 8th June 2010

[15] European Bioinformatics Institute Press Release, "GSK and Online Communities Create Unique Alliance to Stimulate Open Source Drug Discovery for Malaria", http://collaborativedrug.com/blog/news/2010/05/20/gsk-opens-up-2/ Last Accessed on 8th June 2010

[16] Jeff Howe, "The Rise of Crowdsourcing", Wired Magazine, Issue 14.06, June 2006

[17] http://www.ncbi.nlm.nih.gov/pubmed/ Last Accessed on 8th June 2010

8.1 Community Objectives

The majority of ChemSpider's data curation challenges are concerned with the identification of chemical identifiers, adherence to nomenclature structure standards, associated layered information such as experimental data, and establishing dataset record links to publications. With 25 million unique compounds across 300 sources, even simple data imperfections such as spelling errors can quickly make the curation effort unfeasible. Drawing upon experiences with Wikipedia's chemical data curation, ChemSpider engaged its community to assist with curation resources and quality.

8.2 Curation Approach & Types

ChemSpider uses a flat meritocracy model for their curation activities. Normal curators are responsible for deposition which is checked and verified by master curators. Normal curators in turn, can be invited to become masters after some qualifying monitoring period. The curation process is iterative, with normal curators receiving correction comments on rejected structures to apply before any new deposition.

ChemSpider blends human and computer based curation approaches to extract maximum knowledge from its community participants. *Robotic Curation* uses algorithms for error correction and data cleansing at deposition time. The algorithms automatically factor in a higher rank for previous manual edits with master curators occupying the most powerful rank position.

Leveraging novel approaches to curation *Blink* or *Game Based* curation used the gaming paradigm to extract curation effort. A spectral game[18] powered by chemical data from ChemSpider is used as a teaching tool on NMR spectrum interpretation. Game activity is actively monitored to identify problematic issues and paths which are fed back into the curation process to re-check existing spectrum analysis efforts and improve the data set quality. Spectrum analysis also represents *Focused curation* which specifically looks at a particular type of data. Focused curation relies upon targeted expert curators to help specify specific algorithmic and rule development. All curated data tracks specific provenance, including that of change through its deposition parameters.

Recognition of community curation effort and contribution to the wider community as done by Wikipedia is considered as a necessary next step by ChemSpider.

[18] http://www.spectralgame.com/ Last Accessed on 8th June 2010

9 Case Study: Protein Data Bank, Pre-competitive Bioinformatics

The Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB)[19] is dedicated to improving understanding of the function of biological systems through the study of the 3-D structure of biological macromolecules. Started in 1971 with 3 core members it originally offered free access to 7 crystal structures which has grown to the current 63,000 structures available freely online. In 2003, the RCSB PDB joined with sister organizations from Europe (PDBe) and Japan (PDBj) to promote outreach, education and standardization through the wwPDB foundation[20]. Today, the PDB has had over 300 million dataset downloads and continues to curate and annotate data with the support of its community. Its tool and resource offering has grown from a curated data download service, to one that serves complex molecular visualized, search, and analysis capabilities.

9.1 Serving the Community

Community uses of the data are varied; structural biologists and crystallographers starting a project find it useful to check if similar or identical structures are already in place. This can accelerate investigations to the next phase of the experimental process. Protein investigators searching for the 'holy grail' of protein sequence prediction, use the PDB as a knowledge base (assisted by analytics) from which to try and predict structure from sequence. Others use it as a storage and archival mechanism. Pharmaceutical companies frequently download the entire dataset and combine it with proprietary data as an essential tool for their internal drug development.

From PDB's perspective open data sharing is key as it encourages a wider audience in solving the same problem due to basic data and most often pre-competitive data availability. In fact this was the approach taken to combat the Aids virus. Protein - protein interaction and structure comparison represents a field that emerged from having an open structured data set that readily lent itself to structured informatics. PDB invested heavily in community engagement and education to get the point across that progress can be made faster by increasing data availability.

9.2 Curation Approaches & Types

Making available molecular representation, their 3-D coordinates and experimental data requires massive levels of curation to ensure that data inconsistencies are identified and corrected. A central data repository and sister sites accepts data in

[19] http://www.pdb.org/ Last Accessed on 8th June 2010

[20] http://www.wwpdb.org/ Last accessed on 8th June 2010

multiple formats such as the legacy PDB format, the mmCif introduced in 1996 and the current PDBML valid since 2005. Operating a global hierarchical governance approach to data curation work flow, wwPDB staff review and annotating each submitted entry before robotic curation checks for plausibility as part of the data deposition, processing and distribution. Distributing the curation workload across their sister sites helps to manage the activity.

A significant amount of curation process is performed with the use of vocabularies. Vocabularies provide identifier and access support for resource mapping and relationships between biological datasets, varying from organ tissue to structure description. The use of standardized vocabularies also helps with nomenclature used to describe protein and small molecule names and their descriptors present in the structure entry. Nomenclature standardization affects experiment descriptors, information sources and names that also change over time and across different theme lines. Tracking the identifiers across multiple databases even with taxonomic support is a challenge. While large portions of metadata is represented as RDF triples core atomic coordinate data is not. PDB are also involved in the development of experimental science-based ontologies through the Open Biology and Biomedical Ontologies (OBO) project[21] to further assist with the integration effort. The internal curation processes also uses a data dictionary to manage and translate the data into semantic enabled visual representations.

Robotic curation automates data checking with curators contributing to rule definition for inconsistencies. The process is iterative and corrections to discovered mistakes such as those against current standards are applied retrospectively to the archives. Versioned weekly schedules and periodic full datasets are released to keep all sources consistent with the current standards and ensure curation quality.

PDB provenance model was developed prior to the advent of any open public model and a separate project is looking at the Open Provenance Model.

9.3 Observations

Persistence in promoting the open data idea was the single biggest contributor to PDB's success. The PDB engaged community stakeholders from the start to determine what they wanted. Both producers and consumers of the data were engaged in dialogue and outcomes articulated as white papers. This early stakeholder buying, along with patience proved crucial for acceptance. The idea of data sharing was not initially popular but the drive to share important data beyond an individual group has since gained acceptance.

What is of particular interest is that while the wwPDB operate a consortium-type governance structure where consensus decision making is required, its member sites employ a *friendly competitive approach* to the provision of services and software offerings available from their site which they feel better serves the community. Pack-

[21] http://www.obofoundry.org/ Last accessed on 8th June 2010

aging of unique information only available at a particular site "supports innovative use of data and also further helps discovery of representation inconsistencies" observes Helen Berman, Director of PDB, feels that channeling creativity into the web sites is the best way to extract the richness of data usage. The approach of a no cost friendly competitions where you compete with the people you collaborate with "is an example of a lightly regulated community getting what they deserve".

10 Case Study Learnings

Learnings from case study practitioners fell into the broad categories of being either insights or practicalities. Of the two, insight, relating to social best practice was consistently emphasized as the key success factor for project success, and community participation. Practicalities confined to a technical focus were deemed important for project success only. The social elements of community participation along with technical best practices are next discussed.

10.1 Social Best Practices

The successful communities observed had social best practices in common:

- **Participation:** Stakeholders involvement for both data producers and consumers must occur early in the project. This will help provide insight into the basic questions of what they want to do, for whom, and what it will provide. White papers are an effective means to present these ideas, and solicit opinion from the community. They can be used to establish an informal 'social contract' for the community.
- **Engagement:** Outreach activities are essential for promotion and feedback. Social communication and networking forums are useful but be aware that with typical consumers-to-contributors ratios of less than 5%, the majority of your community may not communicate using these media. The communication by email still remains important.
- **Incentives:** For a community to participate in sheer curation there must be a line of sight from the data curating activity, to tangible exploitation benefits. If the general community lacks awareness of the value proposition, a collaborative contribution environment will be slow to emerge. Recognizing contributing curators through a formal feedback mechanism will help reinforce the contribution culture and directly increase output quality.
- **Community Governance Models:** An effective governance structure is vital to ensure the success of a community. Internal communities and consortium perform well when they leverage traditional corporate and democratic governance models. However, these approaches are not appropriate for open communities

where there is a need to engage the community within the governance process. Successful governance models for open communities follow less orthodox approaches using *meritocratic* and *autocratic* principles.

- A *meritocratic* community is lead by an elected leadership team or 'board'. In the meritocratic mode appointments are made and responsibilities assigned to individuals/organizations based upon demonstrated talent and ability. The community operates with an almost completely 'flat' structure, which means that any participant willing to contribute can engage and gain influence in recognition of their contributions. Examples include Apache Software Foundation and ChemSpider.
- An *autocratic* benevolent dictatorship is a community controlled in a hierarchical fashion with a single person or organization leading the community and has final say in decisions. A hierarchical model requires the leader to be strong in diplomacy and community building skills.

10.2 Technical Best Practices

In terms of infrastructure and process support the following were highlighted across the case studies as key practices:

- **Data Representation:** Data representations that are robust and standardized will encourage community usage, and tools development. Support for legacy data formats should be considered, as is the ability to translate all data forward to deal with emergent technology and standards is important.
- **Balancing Human- and Computer-based Curation:** Arriving at a balance between orchestrating automated and human assisted curation will improve data quality. For large datasets robotic curation should be used for validating data deposition and entry, while the community targets focused curation tasks. Robotic curation should always defer to, and never override, human curation edits.
- **Track Provenance:** A user consuming data generated from third parties needs mechanisms to assess the entities and process involved in the generation and publication of this data. Provenance is a key aspect in the process of mapping the historical trail behind an information artifact and can help determining if the data is high quality, trustworthy and compliant. Different users can lead to different perspectives of provenance. A scientist may need to evaluate the fine grained experiment description behind the data, while for a business analyst the 'brand' of data provider can be sufficient for determining quality. The ability to provide a provenance description attached to the data plays an important role in the data quality process. All curation activities including edits, especially where human curators are involved should be recorded and maintained as part of a larger data provenance effort.
- **Data Consumption Infrastructure:** As open datasets become more prevalent, companies will need to develop appropriate internal infrastructures to consume,

curate, manage and integrate third-party data. External data can be generated by business partners, expert communities or from the open web, the organizations data governance policies will need to cater for this consumption.

11 Conclusion

With increased utilization of data within their operational and strategic processes, enterprises need to ensure data quality and accuracy. Data curation is a process that can ensure the quality of data and its fitness for use. Data curation teams have found it difficult to scale the traditional centralized approach and have tapped into community crowd-sourcing and automated and semi-automated curation algorithms.

The emergence of pre-competitive data collaborations is a significant development: pre-competitive data is information that can be shared without conferring a commercial advantage. Within these collaborations, competing organizations share data and the effort required to curate data. With ever increasing data volumes, and continuing pressure on resource availability, these collaborations will become more prevalent within the enterprise information landscape.

Effective community based data curation is highly dependent on the community it serves. Early involvement of key stakeholders, a continuous community communication channel, clear incentives, and an effective governance model are important social aspects for community development. Persistence in promoting the idea of open data is the biggest contributory factor for a successful community.

Acknowledgements In writing this chapter we were fortunate to be have access to a number of thought leaders in the area willing to share their time, insights and experiences. We would like to thank Evan Sandhaus (Semantic Technologist), Rob Larson (Vice President Product Development and Management), and Gregg Fenton (Director Emerging Platforms) from the New York Times, Krista Thomas (Vice President, Marketing & Communications), Tom Tague (OpenCalais initiative Lead) from Thomson Reuters, Antony Williams (VP of Strategic Development) from ChemSpider, Helen Berman (Director), John Westbrook (Product Development) from the Protein Data Bank and finally Nick Lynch (Architect with AstraZeneca) from the Pistoia Alliance. The work presented in this chapter has been funded by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

References

1. Davenport, T.H., Competing On Analytics, in Harvard Business Review. 2006. p. 98-107.
2. Wang, R. and D. Strong, Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 1996. 12(4): p. 5-33.
3. Knight, S.A. and J. Burn, Developing a Framework for Assessing Information Quality on the World Wide Web. Informing Science, 2005. 8: p. 159-172.
4. Ball, A., Preservation and Curation in Institutional Repositories. 2010, Digital Curation Centre.

5. Bourne, P. and J. McEntyre, Biocurators: Contributors to the World of Science. PLoS Comput Biol, 2006. 2(10): p. 142.
6. Uren, V., et al., Semantic Annotation for Knowledge Aanagement: Requirements and a Survey of the State of the Art. Web Semantics: Science, Services and Agents on the World Wide Web, 2006. 4(1): p. 14-28.
7. Appelt, D.E. and D.J. Israel, Introduction to Information Extraction Technology. in International Joint Conference on Artificial Intelligence. 1999.
8. Ekins, S. and A.J. Williams, Reaching out to Collaborators: Crowdsourcing for Pharmaceutical Research. Pharmaceutical Research. 27(3): p. 393-5.
9. Bingham, A. and S. Ekins, Competitive Collaboration in the Pharmaceutical and Biotechnology Industry. Drug Discovery Today, 2009. 14(23-24): p. 1079-81.
10. Barnes, M.R., et al., Lowering Industry Firewalls: Pre-competitive Informatics Initiatives in Drug Discovery. Nature Reviews Drug Discovery, 2009. 8(9): p. 701-708.
11. Giles, J., Internet Encyclopaedias go Head to Head. Nature, 2005. 438(7070): p. 900-901.
12. Emigh, W. and S.C. Herring. Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias. in System Sciences, 2005. HICSS 2005. Proceedings of the 38th Annual Hawaii International Conference on System Sciences.
13. Mons, B., et al., Calling on a Million Minds for Community Annotation in WikiProteins. Genome Biology, 2008. 9(5): R89.
14. Stvilia, B., et al., Information Quality Work Organization in Wikipedia. Journal of the American Society for Information Science and Technology, 2008. 59(6): p. 983-1001.
15. Kollock, P. and M. Smith, The Economies of Online Cooperation: Gifts and Public Goods in Cyberspace, in Communities in Cyberspace. 1999, Routledge. p. 220-239.
16. Bryant, S., A. Forte, and A. Bruckman. Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. in GROUP '05: Proceedings of the 2005 international ACM SIGGROUP Conference on Supporting Group Work. 2005. Sanibel Island, Florida, USA: ACM.
17. Viegas, F., et al. Talk Before You Type: Coordination in Wikipedia. in System Sciences, 2007. HICSS 2007. Proceedings of the 40th Annual Hawaii International Conference on System Sciences.
18. Bizer, C., et al., DBpedia - A Crystallization Point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web, 2009. 7(3): p. 154-165.
19. Hepp, M., K. Siorpaes, and D. Bachlechner, Harvesting Wiki Consensus: Using Wikipedia Entries as Vocabulary for Knowledge Management. IEEE Internet Computing, 2007. 11(5): p. 54-65.

http://www.springer.com/978-1-4419-7664-2

Linking Enterprise Data
(Ed.)D. Wood
2010, XXVI, 291 p., Hardcover
ISBN: 978-1-4419-7664-2