Data Curation Technical Working Group

White Paper

Project Acronym:	**BIG**
Project Title:	Big Data Public Private Forum (BIG)
Project Number:	**318062**
Instrument:	**CSA**
Thematic Priority:	**ICT-2011.4.4**

D2.2.1 First Draft of Technical White Papers (Data Curation)

Work Package:	*WP2 Strategy & Operations*	
Due Date:		28/02/2014
Submission Date:		11/03/2014
Start Date of Project:		01/09/2012
Duration of Project:		26 Months
Organisation Responsible of Deliverable:		NUIG
Version:		1.0
Status:		Final
Author name(s):	Edward Curry Andre Freitas Umair Ul Hassan	NUIG NUIG NUIG
Reviewer(s):	Helen Lippell	PA
Nature:	☒ R – Report ☐ P – Prototype ☐ D – Demonstrator ☐ O - Other	
Dissemination level:	☒ PU - Public ☐ CO - Confidential, only for members of the consortium (including the Commission) ☐ RE - Restricted to a group specified by the consortium (including the Commission Services)	
Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		

Revision history			
Version	**Date**	**Modified by**	**Comments**
0.1	25/04/2013	Andre Freitas, Aftab Iqbal, Umair Ul Hassan, Nur Aini (NUIG)	Finalized the first version of the whitepaper
0.2	27/04/2013	Edward Curry (NUIG)	Review and content modification
0.3	27/04/2013	Helen Lippell (PA)	Review and corrections
0.4	27/04/2013	Andre Freitas, Aftab Iqbal (NUIG)	Fixed corrections
0.5	20/12/2013	Andre Freitas (NUIG)	Major content improvement
0.6	20/02/2014	Andre Freitas (NUIG)	Major content improvement
0.7	15/03/2014	Umair Ul Hassan	Content contribution (human computation, case studies)
0.8	10/03/2014	Helen Lippell (PA)	Review and corrections
0.9	20/03/2014	Edward Curry (NUIG)	Review and content modification
1.0	26/03/2014	Andre Freitas (NUIG)	Final review and minor corrections

Copyright © 2012, BIG Consortium

The BIG Consortium (http://www.big-project.eu/) grants third parties the right to use and distribute all or parts of this document, provided that the BIG project and the document are properly referenced.

THIS DOCUMENT IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS DOCUMENT, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Table of Contents

Index of Figures

Index of Tables

Abbreviations and Acronyms

Abbreviations and Acronyms:

CMS	Content Management System
CbD	Curating by Demonstration
DCC	Digital Curation Centre
ETL	Extract-Transform-Load
ML	Machine Learning
NLP	Natural Language Processing
MDM	Master Data Management
MIRIAM	Minimum Information Required In The Annotation of Models
NGO	Non Governmental Organization
PbD	Programming by Demonstration
PPP	Public Private Partnerships
RDF	Resource Description Framework
SQL	Structured Query Language
W3C	World Wide Web Consortium
XML	Extensible Markup Language

1. Executive Summary

With the emergence of data environments with growing data *variety* and *volume*, organisations need to be supported by processes and technologies that allow them to produce and maintain high quality data facilitating data reuse, accessibility and analysis. In contemporary data management environments, *data curation infrastructures* have a key role in addressing these common challenges found across many different data production and consumption environments. Recent changes in the *scale* of the data landscape bring major changes and new demands to data curation processes and technologies.

This whitepaper investigates how the emerging *Big Data* landscape is defining new requirements for data curation infrastructures and how curation infrastructures are evolving to meet these challenges. The role played by Data Curation is analysed under the context of the Big Data value chain (Figure 1-1). Different dimensions of scaling-up data curation for Big Data are investigated, including emerging technologies, economic models, incentives/social aspects and supporting standards.

This analysis is grounded by literature research, interview with domain experts, surveys and case studies and provide an overview of the state-of-the-art, future requirements and emerging trends in the field.

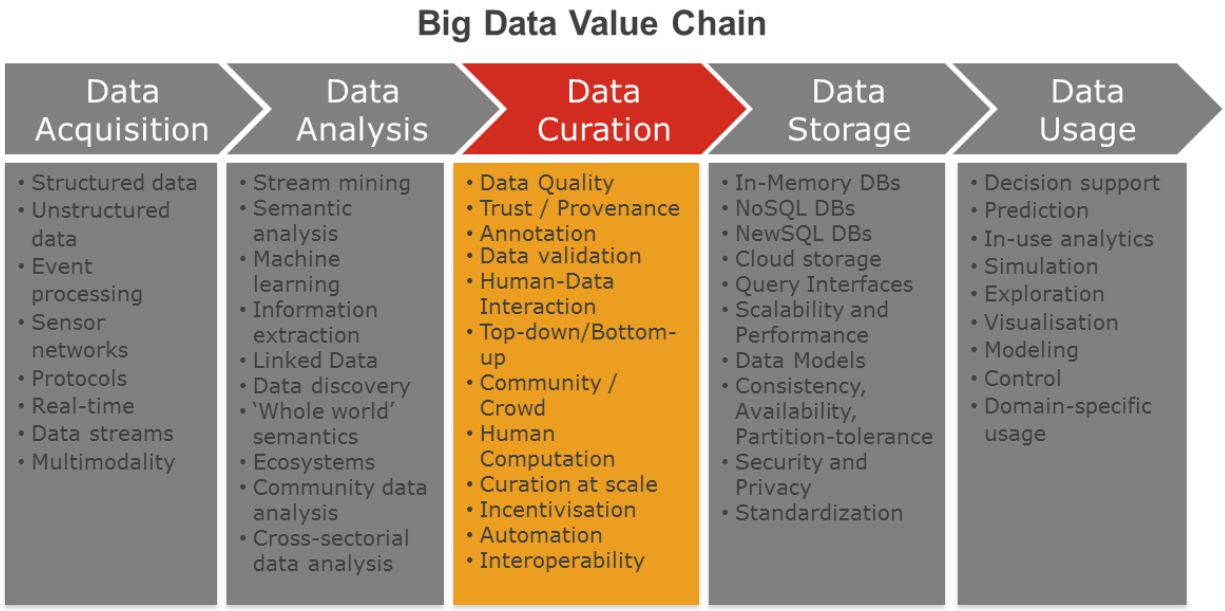

Figure 1-1 Big Data value chain.

2. Big Data Curation Key Insights

The key insights of the data curation technical working group are as follow:

eScience and eGovernment are the innovators while Biomedical and Media companies are the early adopters. The demand for data interoperability and reuse on eScience, and the demand for effective transparency through open data in the context of eGovernment are driving data curation practices and technologies. These sectors play the roles of visionaries and innovators in the data curation technology adoption lifecycle. From the industry perspective, organisations in the biomedical space, such as Pharmaceutical companies, play the role of early-adopters, driven by the need to reduce the time-to-market and lower the costs of the drug discovery pipelines. Media companies are also early adopters, driven by the need to organise large unstructured data collections, to reduce the time to create new products repurposing existing data and to improve accessibility and visibility of information artefacts.

The core impact of data curation is to enable more complete and high quality data-driven models for knowledge organisations. More complete models support a larger number of answers through data analysis. Data curation practices and technologies will *progressively become more present on contemporary data management* environments, facilitating organisations and individuals to *reuse third party data in different contexts*, reducing the barriers for generating content with high data quality. The ability to efficiently cope with data quality and heterogeneity issues at scale will support data consumers on the creation of more sophisticated models, *highly impacting the productivity of knowledge-driven organisations*.

Data curation depends on the creation of an incentives structure. As an emergent activity, there is still vagueness and poor understanding on the role of data curation inside the Big Data lifecycle. In many projects the data curation costs are not estimated or underestimated. The individuation and recognition of the *data curator role* and of data curation activities depends on realistic estimates of the costs associated with producing high quality data. Funding boards can support this process by requiring an explicit estimate of the data curation resources on public funded projects with data deliverables and by requiring the publication of high quality data. Additionally, the improvement of the tracking and recognition of data and infrastructure as a first-class scientific contribution is also a fundamental driver for methodological and technological innovation for data curation and for maximizing the return of investment and reusability of scientific outcomes. Similar recognition is needed within the Enterprise context.

Emerging economic models can support the creation of data curation infrastructures. *Pre-competitive* and *public-private partnerships* are emerging economic models that can support the creation of data curation infrastructures and the generation of high quality data. Additionally, the justification for the investment on data curation infrastructures can be supported by a better quantification of the economic impact of high quality data.

Curation at scale depends on the interplay between automated curation platforms and collaborative approaches leveraging large pools of data curators. Improving the scale of data curation depends on reducing the cost per data curation task and increasing the pool of data curators. *Hybrid human-algorithmic data curation approaches* and the ability to compute the *uncertainty of the results* of algorithmic approaches are fundamental for improving the automation of complex curation tasks. Approaches for automating data curation tasks such as *curation by demonstration* can provide a significant increase in the scale of automation. *Crowdsourcing* also plays an important role in scaling-up data curation, allowing access to large pools of potential data curators. The improvement of crowdsourcing platforms towards more *specialized, automated, reliable* and *sophisticated platforms* and *the improvement of the integration between organizational systems and crowdsourcing platforms* represent an exploitable opportunity in this area.

The improvement of human-data interaction is fundamental for data curation. Improving approaches in which *curators can interact with data* impacts curation efficiency and reduce the barriers for domain experts and casual users to curate data. Examples of key functionalities in human-data interaction include: *natural language interfaces*, *semantic search*, *data summarization & visualization*, and *intuitive data transformation interfaces*.

Data-level trust and permission management mechanisms are fundamental to supporting data management infrastructures for data curation. Provenance management is a key enabler of trust for data curation, providing curators the context to select data that they consider trustworthy and allowing them to capture their data curation decisions. Data curation also depends on mechanisms to assign permissions and digital rights at the data level.

Data and conceptual model standards strongly reduce the data curation effort. A standards-based data representation reduces syntactic and semantic heterogeneity, improving interoperability. Data model and conceptual model standards (e.g. vocabularies and ontologies) are available in different domains. However, their adoption is still growing.

Need for improved theoretical models and methodologies for data curation activities. Theoretical models and methodologies for data curation should concentrate on supporting the transportability of the generated data under different contexts, facilitating the detection of data quality issues and improving the automation of data curation workflows.

Better integration between algorithmic and human computation approaches. The growing maturity of data-driven statistical techniques in fields such as Natural Language Processing (NLP) and Machine Learning (ML) is shifting their use from academic into industry environments. Many NLP and ML tools have uncertainty levels associated with their results and are dependent on training over large training datasets. Better integration between statistical approaches and human computation platforms is essential to allow the continuous evolution of statistical models by the provision of additional training data and also to minimize the impact of errors in the results.

The European Union has a leading role in the technological development of data curation approaches. With visionary and large-scale projects in eScience, and early adopters in the eGovernment and Media sectors, the EU has a leadership role on the technical advance of data curation approaches. The increasing importance of data curation in data management brings a major economic potential for the development of data curation products and services. However, within the EU context data curation technologies still need to be transferred to industry and private-sector initiatives.

3. Introduction

One of the key principles of data analytics is that the quality of the analysis is dependent on the quality of the information analysed. Gartner estimates that more than 25% of critical data in the world's top companies is flawed (Gartner, 2007). Data quality issues can have a significant impact on business operations, especially when it comes to the decision making processes within organisations (Curry et al., 2010).

The *emergence of new platforms for decentralised data creation* such as sensor and mobile platforms, the *increasing availability of open data on the Web* (Howe et al., 2008), added to the *increase in the number of data sources inside organisations* (Brodie & Liu, 2010), brings an unprecedented volume of data to be managed. In addition to the data volume, data consumers in the Big Data era need to cope with *data variety,* as a consequence of the *decentralized data generation*, where data is *created* under different *contexts* and *requirements*. Consuming third-party data comes with the intrinsic cost or repurposing, adapting and ensuring data quality for its new context.

Data curation provides the *methodological* and *technological* data management support to address *data quality issues* maximizing the usability of the data. According to Cragin et al. (2007), *"Data curation is the active and on-going management of data through its lifecycle of interest and usefulness; … curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time"*. Data curation emerges as a *key data management process* where there is an increase in the number of data sources and platforms for data generation.

Data curation processes can be categorized into different activities such as *content creation*, *selection*, *classification*, *transformation*, *validation* and *preservation*. The selection and implementation of a data curation process is a multi-dimensional problem, depending on the interaction between the *incentives*, *economics*, *standards* and *technologies* dimensions. This whitepaper analyses the data dynamics in which data curation is inserted, provides the description of key concepts for data curation activities and investigates the future requirements for data curation.

3.1. Emerging Requirements for Big Data: Variety & Reuse

Many Big Data scenarios are associated to reusing and integrating data from a number of different data sources. This perception is recurrent across data curation experts and practitioners and it is reflected on statements such: "*a lot of Big Data is a lot of small data put together*", "*most of Big Data is not a uniform big block*", "*each data piece is very small and very messy, and a lot of what we are doing there is dealing with that variety*" (Data Curation Interview: Paul Groth, 2013).

Reusing data which was generated under different requirements comes with the intrinsic price of coping with *data quality* and *data heterogeneity* issues. Data can be incomplete or may need to be transformed in order to be rendered useful. Kevin Ashley, director of Digital Curation Centre summarizes the mindset behind data reuse: *"… [it is] when you simply use what is there, which may not be what you would have collected in an ideal world, but you may be able to derive some useful knowledge from it*" (Data Curation Interview: Kevin Ashley, 2013). In this context, data shifts from a resource that is tailored from the start to a certain purpose, to a raw material that will need to be repurposed in different contexts in order to satisfy a particular requirement.

In this scenario *data curation* emerges as a key data management activity. Data curation can be seen from a *data generation* perspective (curation at source), where data is represented in a way that maximizes its quality in different contexts. Experts emphasize this as an important aspect of data curation: from the *data science* aspect, we need methodologies to describe data so that it is actually reusable outside its original context (Data Curation Interview: Kevin Ashley, 2013). This points to the demand to investigate approaches which maximize the quality of the data in multiple contexts with a minimum curation effort: "*we are going to curate data in a way that makes it usable ideally for any question that somebody might try to ask the data*" (Data Curation Interview: Kevin Ashley, 2013). Data curation can also be done at the *data consumption* side where data resources are selected and transformed to fit a set of requirements from the data consumption side.

Data curation activities are heavily dependent on the challenges of scale, in particular data variety that emerges in the Big Data context. James Cheney, research fellow at the University of Edinburgh, observes that "*Big Data seems to be about addressing challenges of scale, in terms of how fast things are coming out at you versus how much it costs to get value out of what you already have*". Additionally, "*if the programming effort per amount of high quality data is really high, the data is big in the sense of high cost to produce new information*" (Data Curation Interview: James Cheney, 2013). Coping with data variety can be costly even for smaller amounts of data: "*you can have Big Data challenges not only because you have Petabytes of data but because data is incredibly varied and therefore consumes a lot of resources to make sense of it*". James Cheney complements: "*if the amount of money that you need to spend at data cleaning is doubling every year, even if you are only dealing with a couple of MBs that's still a Big Data problem*".

While in the Big Data context the expression *data variety* is used to express the data management trend of coping with data from different sources, the concepts of *data quality* (Wang & Strong, 1996; Knight & Burn, 2005) and *data heterogeneity* (Sheth, 1999) have been well established in the database literature and provide a precise ground for understanding the tasks involved in data curation.

3.2. Emerging Trends: Scaling-up Data Curation

Despite the fact that data heterogeneity and data quality were concerns already present before the Big Data scale era, they become more prevalent in data management tasks with the *growth in the number of data sources*. This growth brought the need to define *principles* and *scalable approaches* for *coping with data quality issues*. It also brought data curation from a niche activity, restricted to a small community of scientists and analysts with high data quality standards, to a routine data management activity, which will progressively become more present within the average data management environment.

The growth in the number of data sources and the scope of databases defines a *long tail of data variety*. Traditional relational data management environments were focused on data which mapped to frequent business processes and were regular enough to fit into a relational model. The *long tail of data variety* (Figure 3-1) expresses the shift towards expanding the data coverage of data management environments towards data which is less frequently used, more decentralized, and less structured. The long tail allows data consumers to have a more comprehensive model of their domain that can be *searched*, *queried*, *analysed* and *navigated*.

The central challenge of data curation models in the Big Data era is to deal with the long tail of data and to *improve data curation scalability*, by *reducing the cost* of data curation and *increasing the number of data curators* (Figure 3-1), allowing data curation tasks to be addressed under limited time constraints.

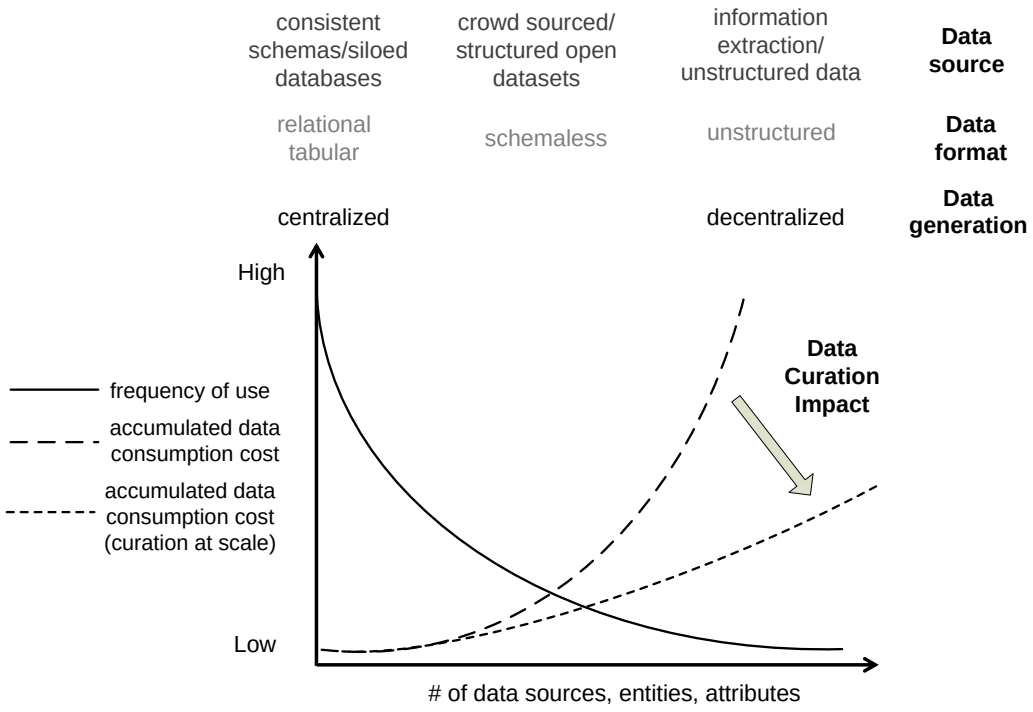

Figure 3-1 The long tail of data curation and the scalability of data curation activities.

Scaling up data curation is a multidisciplinary problem that requires the development of *economic models*, *social/incentive structures* and *standards*, in coordination with *technological solutions*. The connection between these dimensions and data curation scalability is at the centre of the future requirements and future trends for data curation.

4. Social & Economic Impact

The growing availability of data brings the opportunity for people to use them to inform their decision making process, allowing data consumers to have a more complete *data-supported* picture of the reality. While some Big Data use cases are based on large scale but small schema, regular datasets, other decision-making scenarios depend on the integration of complex, multi-domain and distributed data. The extraction of value from information coming from different data sources is dependent on the feasibility of integrating and analysing these data sources.

Decision makers can range from molecular biologists to government officials or marketing professionals and they have in common the need to discover patterns and create models to address a specific task or business objective. These models need to be supported by quantitative evidence. While unstructured data (such as text resources) can support the decision making process, *structured data provides to users greater analytical capabilities, by defining a structured representation associated with the data*. This allows users to compare, aggregate and transform data. With more data available, the barrier of data acquisition is reduced. However, to extract value from it, data needs to be systematically processed, transformed and repurposed into a new context.

Areas which depend on the representation of multi-domain and complex models are leading the data curation technology lifecycle. eScience projects lead the experimentation and innovation on data curation, and are driven by the need of creating infrastructures for improving reproducibility and large-scale multidisciplinary collaboration in science. They play the role of *visionaries* in the *technology adoption lifecycle for advanced data curation technologies* (see Use Cases Section).

On *the early adopter* phase of the lifecycle, the biomedical industry (in particular, the Pharmaceutical industry) is the main player, *driven by the need of reducing the costs and time-to-market of drug discovery pipelines* (Data Curation Interview: Nick Lynch, 2013). For Pharmaceutical companies data curation is central to organisational data management and third-party data integration. Following a different set of requirements, *the media industry is also positioned as early-adopters*, using data curation pipelines to classify large collections of unstructured resources (text and video), improving the data consumption experience through better accessibility and maximizing its reuse under different contexts. *The third major early adopters are governments*, targeting transparency through open data projects (Shadbolt et al., 2012).

Data curation enables the extraction of value from data, and it is a capability that is required for areas that are dependent on complex and/or continuous data integration and classification. The

Future Use Case Scenario

A government data scientist wants to understand the distribution of costs in hospitals around the country. In order to do that she needs to collect and integrate data from different data sources. She starts using a government search engine where all public data is catalogued. She quickly finds the data sources that she wants and collects all the data related to hospital expenses. Different hospitals are using different attributes to describe the data (e.g., item vs material, cost vs price). The curation application detects the semantic similarities and automatically normalizes the vocabularies of the data sources. For the normalization, the curation application interacts with the user to get some confirmation feedbacks. The user is able to analyse the hospitals that deviate from the cost average. To provide further context, the user searches for reference healthcare data in other countries. For this she searches open data in an open data catalogue search engine, quickly integrating with her existing data. To do the integration, language translation and currency conversion services are automatically invoked. With this data at hands she selects the hospitals with higher costs and analyses the distribution of cost items. For some of the data items, the curation platform signals a data quality error message, based on the provenance metadata. In order to correct this error, a report is used, where the correct data is extracted from the report text into the curated dataset. The analyst discovers the anomalous cost items. While preparing to publish the results, the data curation platform notifies that some actions over the data will improve searchability and will facilitate data integration. After doing the modifications, she publishes the analyses on a report that is directly linked to the curated dataset. The data curation platform automatically compiles the descriptors for the dataset and records it in the public data catalogues. As all the changes in the new dataset are recorded as a provenance workflow, the same analysis can be reproduced and reused by other analyst in the future.

improvement of data curation tools and methods directly provides greater efficiency of the knowledge discovery process, maximize return of Investment per data item through reuse and improve organisational transparency.

5. Core Concepts & State-of-the-Art

5.1. Introduction

Data curation is recently evolving under the demands to manage data which grows in volume and variety, a trend that has intensified over the last few years. Despite the growth in the amount of organisations and practitioners involved in data curation, the field is still under formation and it is highly dynamic. This section introduces a high-level overview of the key concepts related to data curation and briefly depicts the state-of-the-art in this area.

The starting point for any data curation activity is the identification of the use case for creating a curated dataset. Typically, a curation effort will have a number of associated motivations, including improving accessibility, data quality or repurposing data to a specific use. Once the goal is clearly established, one can start to define the curation process. There is no single process to curate data and there are many ways to setup a data curation effort.

The major factors influencing the design of a curation approach include:

- Quantity of data to be curated (including new and legacy data)
- Rate of change of the data
- Amount of effort required to curate the data
- Availability of experts

These factors determine the amount of the work required to curate a dataset. Big Data environments largely impact the major factors of the curation process. While dealing with an infrequently changing and small quantity of data (<1,000 records), with a minimal curation effort per record (minutes), 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 (Curry et al., 2010). Curation groups can deal with large curation efforts, but there is a limit to their scalability. When curating large quantities of dynamic data (>million records) even the most sophisticated and specialized 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 *algorithmic curation approaches*. Different curation approaches are not mutually exclusive and can be composed in different ways. These blended approaches are proving to be successful in existing projects (Curry et al., 2010).

5.2. Lifecycle Model

The core data curation workflow can be categorised into the following elements (see Figure 5-1).

- **Raw data/Curated data:** The data being curated have different characteristics that highly impact the data curation requirements. Data can vary in terms of structure level (unstructured, semi-structured, structured), data model (Relational, XML, RDF, text, etc.), dynamicity, volume, distribution (distributed, centralised), and heterogeneity.
- **Data curators:** Curation activities can be carried out by individuals, organisations, communities, etc. Data curators can have different roles according to the curation activities that they are involved in the data curation workflow.
- **Artefacts, tools, and processes needed to support the curation process:** A number of artefacts, tools, and processes can support data curation efforts, including workflow support, web-based community collaboration platforms, taxonomies, etc. Algorithms can help to automate or semi-automate curation activities such as data cleansing, record duplication and classification algorithms (Curry et al., 2010).

- **Data curation workflow:** Defines how the data curation activities are composed and executed. Different methods for organising the data curation effort can be defined such as through the creation of a curation group/department or a through a sheer curation workflow that enlists the support of users.

Data curation activities can be organised under a lifecycle model. The DCC Curation Lifecycle Model (Higgins, 2008) and the SURF Foundation Lifecycle model (Verhaar, et al., 2010) are examples of data curation lifecycle models. These models were merged and synthesized in Figure 5-1, which covers the main curation activities and their associated environments and actors from data creation to long term preservation.

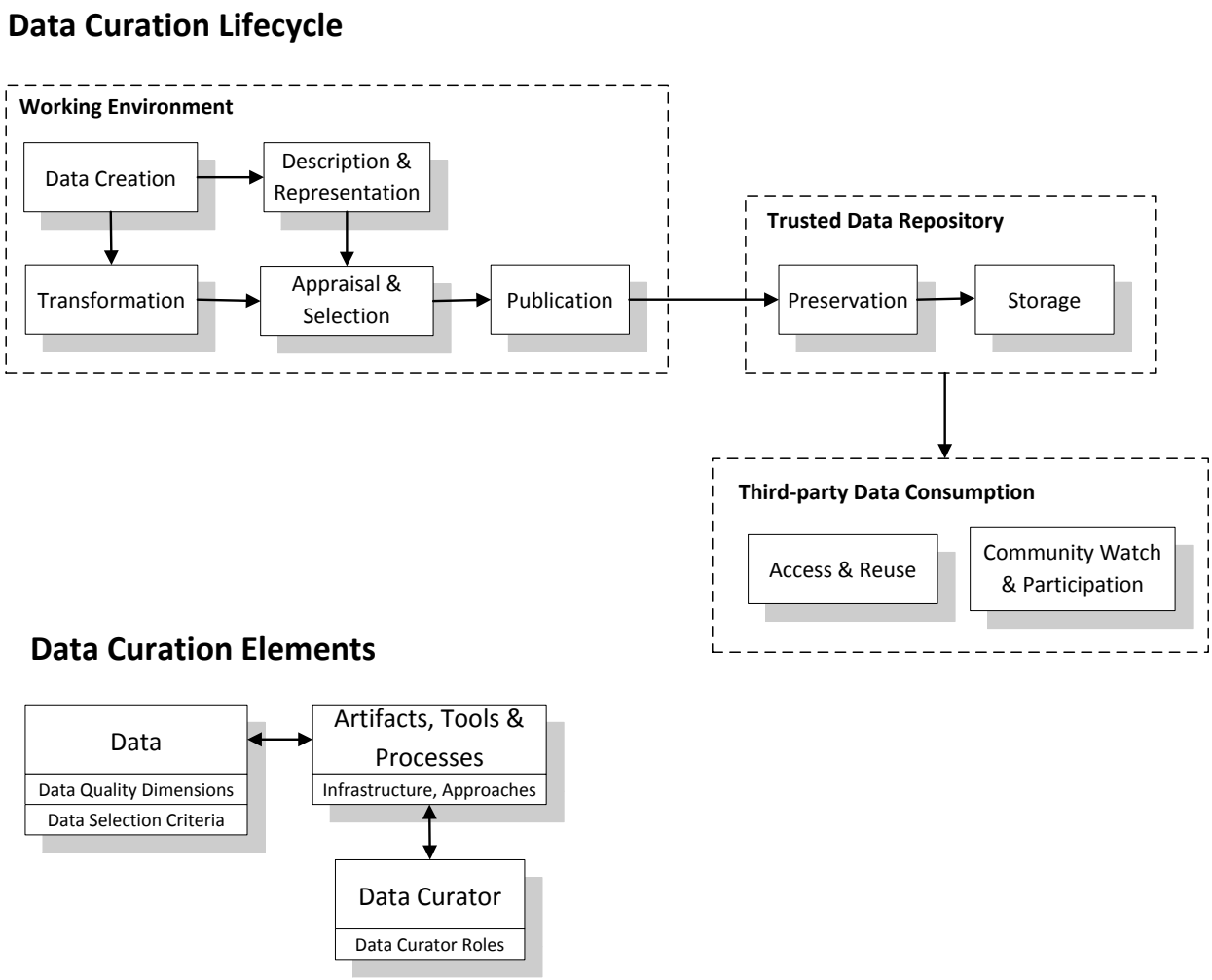

Figure 5-1: The data curation lifecycle based on the DCC Curation Lifecycle Model[1] and on the SURF foundation Curation Lifecycle Model.

The following sections describe and organize the core data curation elements (Figure 5-1).

1 http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf

5.3. Data Selection Criteria

With the growth of digital data in recent years, it is necessary to determine which data is necessary to be kept in long retention due to the large associated data maintenance costs (Beagrie et al., 2008). As such it is necessary to define clear criteria for which data should be curated. With respect to data appraisal, (Eastwood, 2004) described four core activities to appraising digital data: (1) compiling and analysing information, (2) assessing value, (3) determining the feasibility of preservation and (4) making the appraisal decision.

Complementarily to the data appraisal activities, the Data Curation Centre (DCC) introduced a list of indicators to help in the quantified evaluation of data appraisal (Ball, 2010): (1) quantity, (2) timeframe, (3) key records, (4) ownership, (5) requirement to keep data, (6) lifespan of the data, (7) documentation and metadata, (8) technical aspects, (9) economic concerns, (10) access, (11) use and reuse. A subset of these indicators can be used to depict the characteristics of the data which is being curated. A consolidated set of data appraisal indicators was selected from the original DCC list for being more representative for the appraisal process of the selected data curation projects.

5.4. Data Quality Dimensions

The increased utilisation of data, with a wide range of key organisational activities has created a data intensive landscape and caused a drastic increase in the sophistication of data infrastructures in organisations. One of the key principles of data analytics is that the quality of the analysis is dependent on the quality of the information analysed. However, within operational data driven systems, there is a large variance in the quality of information.

Perception of data quality is highly dependent on the fitness for use (Curry et al., 2010); being relative to the specific task that a user has at hand. Data quality is usually described in the scientific literature (Wang & Strong, 1996; Knight & Burn, 2005) by a series of quality dimensions that represent a set of consistency properties for a data artefact. The following data quality dimensions are based on the data quality classification for data curation as proposed in (Curry et al., 2010):

- **Discoverability, Accessibility & Availability:** Addresses if users can find the data that satisfies their information needs and then access it in a simple manner. Data curation impacts the accessibility of the data by representing, classifying and storing it in a consistent manner.
- **Completeness:** Addresses if all the information required for a certain task, including the contextual description, is present in the dataset. Data curation can be used for the verification of omissions of values or records in the data. Curation can also be used to provide the wider context of data by linking/connecting related datasets.
- **Interpretability and Reusability:** Ensures the common interpretation of data. Humans can have different underlying assumptions around a subject that can significantly affect the way they interpret data. Data curation tasks related to this dimension includes the removal of ambiguity, the normalisation of the terminology, and the explicitation of the semantic assumptions.
- **Accuracy:** Ensures that the data correctly represent the “real-world” values it models. Data curation can be used to provide additional levels of verification of the data.
- **Consistency & Integrity:** Covers the uniformity and the semantic consistency of the data under a specific conceptual, representation model and format. Inconsistent data can introduce significant barriers for organizations attempting to integrate different systems and applications. Data curation can be used to ensure that data is consistently created and maintained under standardized terminologies and identifiers.

- **Trustworthiness:** Ensures the reliability in the fact that it is expressed in the data. Data curation tasks can support data consumers in getting explicit reliability indicators for the data. Answers to questions such as: *where did the data come from?*; *which are the activities behind the data?*; *can it be reliably traced back to the original source?*; *what is the reputation of the data sources?*. The provision of a provenance representation associated with can be used to assess the trustworthiness behind the data production and delivery. Data curation activities could be used to determine the reputation of data sources.
- **Timeliness:** Ensures that the information is up-to-date. Data curation can be used to support the classification of the temporal aspect of the data, with respect to the task at hand.

5.5. Data Curation Roles

Human actors play an important role in the data curation lifecycle. Data curation projects evolved in the direction of specifying different roles for data curators according to their associated activity in the data curation workflow. The following categories summarize the core roles for data curators:

- **Coordinator:** Coordinates and manages the data curation workflow.
- **Rules & Policies Manager:** Determines the set of requirements associated with the data curation activities and provide policies and good-practices to enforce the requirements.
- **Schema/Taxonomy/Ontology Manager:** Coordinates the maintenance of the conceptual model and metadata associated with the data.
- **Data Validator:** Validates and oversees specific data curation activities and ensures that the policies and good-practices are being followed by the data curators. This role is also commonly performed by algorithmic approaches.
- **Domain Experts:** Experts in the data curation domain who, in most of the cases, concentrates the core data curation tasks.
- **Data Consumers:** The consumers of the data that can in some cases help in the data curation activities.

5.6. Current Approaches for Data Curation

This section concentrates on describing the technologies that are widely adopted and established approaches for data curation, while the next section focuses on the emerging approaches.

Master Data Management (MDM) are the processes and tools that support a single point of reference for the data of an organization, an authoritative data source. MDM tools can be used to remove duplicates, standardize data syntax and as an authoritative source of master data. Master Data Management (MDM) focuses on ensuring that an organization does not use multiple and inconsistent versions of the same master data in different parts of its systems. Processes in MDM include source identification, data transformation, normalization, rule administration, error detection and correction, data consolidation, data storage, classification, taxonomy services, schema mapping, semantic enrichment.

Master data management (MDM) is highly associated with data quality. According to Morris & Vesset, (2005) the three main objectives of MDM are:

1. Synchronizing master data across multiple instances of an enterprise application
2. Coordinating master data management during an application migration
3. Compliance and performance management reporting across multiple analytic system

Rowe (2012) provides an analysis on how 163 organizations implement MDM and its business impact.

Collaboration Spaces such as Wiki platforms and Content Management Systems (CMSs) allow users to collaboratively create and curate unstructured and structured data. While CMSs focuses on allowing smaller and more restricted groups to collaboratively edit and publish online content (such as News, blogs and eCommerce plaforms), Wikis have proven to scale to very large user bases. As of 2014, Wikipedia, for example counted more than 4,000,000 articles and has a community with more than 130,000 active registered contributors.

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. Important cases include Intellipedia, a deployment of the MediaWiki platform covering 16 U.S. Intelligence agencies, and Wiki Proteins, a collaborative environment for knowledge discovery and annotation (Mons, 2008).

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. All users are allowed to edit Wikipedia contents. Administrators, however, have additional permissions in the system. (Curry et. al, 2010). Most of Wikis and CMS platforms target unstructured and semi-structured data content, allowing users to classify and interlink unstructured content.

Crowdsourcing based on the notion of “wisdom of crowds” advocates that potentially large groups of non-experts can solve complex problems usually considered being solvable only by experts (Surowiecki, 2005). Crowdsourcing has emerged as a powerful paradigm for outsourcing work at scale with the help of online people (Doan et al., 2011). Crowdsourcing has been fuelled by the rapid development in web technologies that facilitate contributions from millions of online users. The underlying assumption is that large-scale and cheap labour can be acquired on the Web. The effectiveness of crowdsourcing has been demonstrated through websites like Wikipedia[1], Amazon Mechanical Turk[2], and Kaggle[3]. Wikipedia follows a volunteer crowdsourcing approach were the general public is asked to contribute to the encyclopaedia creation project. Amazon Mechanical Turk provides a labour market for paid crowdsourcing tasks. Kaggle enables organization to publish problems to be solved through a competition between participants against a predefined reward. Although different in terms of incentive models, all these websites allow access to large number of groups for problem solving. Therefore, enabling their use as recruitment platforms for human computation.

[1] "Wikipedia." 2005. 12 Feb. 2014 <https://www.wikipedia.org/>

[2] "Amazon Mechanical Turk." 2007. 12 Feb. 2014 <https://www.mturk.com/>

[3] "Kaggle: Go from Big Data to Big Analytics." 2005. 12 Feb. 2014 <http://www.kaggle.com/>

6. Future Requirements and Emerging Trends for Big Data Curation

6.1. Introduction

This section aims at providing a *roadmap* for data curation based on a *set of future requirements for data curation* and *emerging data curation approaches* for coping with the requirements. Both future requirements and the emerging approaches were collected by an extensive analysis of the state-of-the-art approaches as described in the methodology (see the Methodology section).

6.2. Future Requirements

This section analyses categories of requirements which are recurrent across state-of-the-art systems and which emerged in the domain expert interviews as a fundamental direction for the future of data curation. The list of requirements was compiled by selecting and categorizing the most recurrent demands in the state-of-the-art survey. Each requirement is categorized according to the following attributes (Table 6-1):

- **Core Requirement Dimensions:** Consists of the main categories needed to address the requirement. The dimensions are: technical, social, incentive, methodological, standardization, economic, and policy.
- **Impact-level:** Consists of the impact of the requirement for the data curation field. By its construction, only requirements above a certain impact threshold are listed. Possible values are: medium, medium-high, high, very high.
- **Affected areas:** Lists the areas which are most impacted by the requirement. Possible values are: Science, Government, Industry Sectors (Financial, Health, Media & Entertainment, Telco, Manufacturing), and Environmental.
- **Priority:** Covers the level of priority that is associated with the requirement. Possible values are: short-term (covers requirements which highly impact further developments in the field, i.e. they are foundational, < 3 years), medium-term (3-7 years) and consolidation (> 7 years).
- **Core Actors:** Covers the main actors that should be responsible for addressing the core requirement. Core actors are: Government, Industry, Academia, NGOs, and User communities.

Requirement Category	Requirement	Core Requirement Dimension	Impact-level	Affected Areas	Priority	Core Actors
Incentives Creation	Creation of incentives mechanisms for the maintenance and publication of curated datasets.	Economic, Social, Policy	Very High	Science, Government, Environmental, Financial, Health	Short-term	Government
Economic Models	Definition of models for the data economy	Economic, Policy	Very High	All sectors	Short-term	Government, Industry

Social Engagement Mechanisms	Understanding of social engagement mechanisms	Social, Technical	Medium	Science, Government, Environmental	Long-term	Academia, NGOs, Industry
Curation at Scale	Reduction of the cost associated with the data curation task (scalability)	Technical, Social, Economic	Very High	All sectors	Medium-term	Academia, Industry, User communities
Human-Data Interaction	Improvement of the Human-Data interaction aspects. Enabling domain experts and casual users to query, explore, transform and curate data.	Technical	Very High	All sectors	Long-term	Academia, Industry
Trust	Inclusion of trustworthiness mechanisms in data curation	Technical	High	All sectors	Short-term	Academia, Industry
Standardization & Interoperability	Integration and interoperability between data curation platforms / Standardization	Technical, Social, Policy, Methodological	Very High	All sectors	Short-term	User communities, Industry, Academia
Curation Models	Investigation of theoretical and domain specific models for data curation	Technical, Methodological	Medium-High	All sectors	Long-term	Academia
Unstructured-Structured Integration	Better integration between unstructured and structured data and tools	Technical	Medium	Science, Media, Health, Financial, Government	Long-term	Academia, Industry

Table 6-1 Future requirements for data curation.

7. Emerging Paradigms

In the state-of-the-art analysis, key social, technical and methodological approaches emerged for addressing the future requirements. In this section, these emerging approaches are described as well as their coverage in relation to the category of requirements. Emerging approaches are defined as approaches that have a limited adoption.

7.1. Incentives & Social Engagement Mechanisms

Open and interoperable data policies: From an incentives perspective the demand for high quality data is the driver of the evolution of data curation platforms. The effort to produce and maintain high quality data needs to be supported by a solid incentives system, which at this point in time is not fully in place. High quality open data can be one of the drivers of societal impact by supporting more efficient and reproducible science (eScience) (Norris, 2007) and more transparent and efficient governments (eGovernment) (Shadbolt et al., 2012). These sectors play the *innovators* and *early adopters* roles in the *data curation technology adoption lifecycle* and are the main drivers of innovation in data curation tools and methods. Funding agencies and policy makers have a fundamental role in this process and should direct and support scientists and government officials in the direction making available their data products in an interoperable way. The demand for high quality and interoperable data can drive the evolution of data curation methods and tools.

Attribution and recognition of data and infrastructure contributions: From the eScience perspective, scientific and editorial committees of prestigious publications have the power to change the methodological landscape of scholarly communication, by emphasizing reproducibility in the review process and by requiring publications to be supported by high quality data when applicable. From the scientist perspective, publications supported by data can facilitate reproducibility and avoid rework and as a consequence, increase scientific efficiency and impact of the scientific products. Additionally, as data becomes more prevalent as a primary scientific product it becomes a citable resource. Mechanisms such as ORCID (Thomson Reuters Technical Report, 2013) and Altmetrics (Priem et al., 2010) already provide the supporting elements for identifying, attributing and quantifying impact outputs such as datasets and software. The recognition of data and software contributions in academic evaluation systems is a critical element for driving high quality scientific data.

Better recognition of the data curation role: The cost of publishing high quality data is not negligible and should be an explicit part in the estimated costs of a project with a data deliverable. Additionally, the methodological impact of data curation requires that the role of the data curator to be better recognised across the scientific and publishing pipeline. Some organisations and projects have already a clear definition of different data curator roles (Wikipedia, NYT, PDB, Chemspider) (see Case Studies Section). The reader is referred to the case studies, to understand the activities of different data curation roles.

Better understanding of social engagement mechanisms: While part of the incentives structure may be triggered by public policies, or by direct financial gain, others may emerge from the direct benefits of being part of a project which is meaningful for a user community. Projects such as Wikipedia[1], GalaxyZoo (Forston et al., 2011) or FoldIt (Khatib et al., 2011) have collected large bases of volunteer data curators exploring different sets of incentive mechanisms, which can be based on visibility & social or professional status, social impact, meaningfulness or fun. The understanding of the principles and the development of the

[1] http://www.wikipedia.org/

mechanisms behind the engagement of large user bases is an important issue for amplifying data curation efforts.

7.2. Economic Models

Currently there are emerging economic models that can provide the financial basis to support the generation and maintenance of high quality data and the associated data curation infrastructures.

Pre-competitive partnerships for data curation: A *pre-competitive collaboration* scheme is one economic model in which a consortium of organisations which are typically competitors collaborate in parts of the Research & Development (R&D) process which does not impact in their commercial competitive advantage. This allows partners to share the *costs* and *risks* associated with parts of the R&D process. One case of this model is the Pistoia Alliance (Wise, 2012), which is a precompetitive alliance of life science companies, vendors, publishers, and academic groups that aims to lower barriers to innovation by improving the interoperability of R&D business processes. Examples of shared resources include data and data infrastructure tools. Pistoia Alliance was founded by Pharmaceutical companies such as AstraZeneca, GSK, Pfizer and Novartis.

Public-private data partnerships for curation: Another emerging economic model for data curation are *public-private partnerships* (PPP), in which private companies and the public sector collaborate towards a mutual benefit partnership. In a PPP the risks, costs and benefits are shared among the partners, which have non-competing, complementary interests over the data. GeoConnections Canada is an example of a national federal/provincial/territorial PPP initiative launched in 1999, with the objective of developing the Canadian Geospatial Data Infrastructure (CGDI) and publishing geospatial information on the Web (Harper, 2012; Data Curation Interview: Joe Sewash, 2013). GeoConnections has been developed on a collaborative model involving the participation of federal, provincial and territorial agencies, and the private and academic sectors. Geospatial data and its high impact for both the public (environmental, administration) and private (natural resources companies) sectors is one of the early cases of PPPs.

Quantification of the economic impact of data: The development of approaches to quantify the economic impact, value creation and associated costs behind data resources is a fundamental element for justifying private and public investments in data infrastructures. One exemplar case of value quantification is the JISC study *"Data centres: their use, value and impact"* (JISC Report, 2011) which provides a quantitative account of the value creation process of eight data centres. The creation of quantitative financial measures can provide the evidential support for data infrastructure investments both public and private, creating sustainable business models grounded on data assets, expanding the existing data economy.

7.3. Curation at Scale

Human computation & Crowdsourcing services: Data curation can be a resource-intensive and complex task, which can easily exceed the capacity of a single individual. Most non-trivial data curation efforts are dependent of a collective data curation set-up, where participants are able to share the costs, risks and technical challenges. Depending on the domain, data scale and type of curation activity, data curation efforts can utilize relevant communities through invitation or crowds (Doan et al., 2011). These systems can range from systems with large and open participation base such as Wikipedia (crowds-based), to systems or more restricted domain expert groups, such as Chemspider.

The notion of "wisdom of crowds" advocates that potentially large groups of non-experts can solve complex problems usually considered to be solvable only by experts (Surowiecki, 2005). Crowdsourcing has emerged as a powerful paradigm for outsourcing work at scale with the help of online people (Doan et al., 2011). Crowdsourcing has been fuelled by the rapid development in web technologies that facilitate contributions from millions of online users. The underlying assumption is that large-scale and cheap labour can be acquired on the Web. The effectiveness of crowdsourcing has been demonstrated through websites like Wikipedia[1], Amazon Mechanical Turk[2], and Kaggle[3]. Wikipedia follows a volunteer crowdsourcing approach were general public is asked to contribute to the encyclopaedia creation project for the benefit of everyone (Kittur et al., 2007). Amazon Mechanical Turk provides a labour market for crowdsourcing tasks against money (Ipeirotis, 2010). Kaggle enables organization to publish problems to be solved through a competition between participants against a predefined reward. Although different in terms of incentive models, all these websites allow access to large numbers of workers, therefore, enabling their use as recruitment platforms for human computation (Law & von Ahn, 2011).

General-purpose crowdsourcing service platforms such as CrowdFlower (CrowdFlower Whitepaper, 2012) or Amazon Mechanical Turk (Iperiotis, 2010) allow projects to route tasks for a paid crowd. The user of the service is abstracted from the effort of gathering the crowd, and offers its tasks for a price in a market of crowd-workers. Crowdsourcing service platforms provide a flexible model, and can be used to address ad hoc small scale-data curation tasks (such as a simple classification of thousands of images for a research project), peak data curation volumes (e.g. mapping and translating data in an emergency response situation) or at regular curation volumes (e.g. continuous data curation for a company). Crowdsourcing service platforms are rapidly evolving but there is still a major space for *market differentiation* and *growth*. CrowdFlower for example is evolving in the direction of *providing better APIs*, supporting *better integration with external systems*.

At crowdsourcing platforms, people show variability in the quality of work they produce, as well as the amount of time they take for the same work. Additionally the accuracy and latency of human processors is not uniform over time. Therefore appropriate methods are required to route tasks to the right person at the right time (Ul Hassan et al, 2012). Furthermore combining work by different people on same task might also help in improving the quality of work (Law & von Ahn, 2009). Recruitment of right humans for computation is a major challenge of human computation.

[1] "Wikipedia." 2005. 12 Feb. 2014 <https://www.wikipedia.org/>

[2] "Amazon Mechanical Turk." 2007. 12 Feb. 2014 <https://www.mturk.com/>

[3] "Kaggle: Go from Big Data to Big Analytics." 2005. 12 Feb. 2014 <http://www.kaggle.com/>

Today, these platforms are mostly restricted to tasks that can be delegated to a paid generic audience. Possible future differentiation avenues include: (i) support for highly specialised domain experts, (ii) more flexibility in the selection of demographic profiles, (iii) creation of longer term (more persistent) relationships with teams of workers, (iv) creation of a major general purpose open crowdsourcing service platform for voluntary work and (v) using historical data to provide more productivity and automation for data curators (Kittur et al., 2013).

Instrumenting popular applications for data curation: In most cases data curation is performed with common office applications: regular spreadsheets, text editors and email (Data Curation Interview: James Cheney, 2013). These tools are an intrinsic part of existing data curation infrastructures and users are familiarized with them. These tools, however, lack some of the functionalities which are fundamental for data curation: (i) capture and representation of user actions; (ii) annotation mechanisms/vocabulary reuse; (iii) ability to handle large-scale data; (iv) better search capabilities; (v) integration with multiple data sources.

Extending applications with large user bases for data curation provide an opportunity for a low barrier penetration of data curation functionalities into more ad hoc data curation infrastructures. This allows wiring fundamental data curation processes into existing routine activities without a major disruption of the user working process (Data Curation Interview: Carole Goble, 2013).

General-purpose data curation pipelines: While the adaptation and instrumentation of regular tools can provide a low cost generic data curation solution, many projects will demand the use of tools designed from the start to support more sophisticated data curation activities. The development of *general-purpose data curation frameworks* that integrate main data curation functionalities to a large-scale data curation platform is a fundamental element for organisations which do large-scale data curation. Platforms such as Open Refine[1] and Karma (Gil et al., 2011), provide examples of emerging data curation frameworks, with a focus on data transformation and integration. Differently from Extract Transform Load (ETL) frameworks, data curation platforms provide a better support for ad hoc, dynamic, manual, less frequent (long tail) and less scripted data transformations and integration, while ETL pipelines can be seen as concentrating recurrent activities which gets more formalized into a scripted process. General-purpose data curation platforms should target domain experts, trying to provide tools that are usable for people outside the Computer Science/Information Technology background.

Algorithmic validation/annotation: Most of the points raised so far in this section are related to expanding the base of curators, lowering the barriers to do curation. Another major direction for reducing the cost of data curation is related to the automation of data curation activities. Algorithms are becoming more intelligent with advances in machine learning and artificial intelligence. It is expected that machine intelligence will be able to validate, repair, and annotate data within seconds, which might take hours for humans to perform (Kong et al., 2011). In effect, humans will be involved as required e.g. for defining curation rules, validating hard instances, or providing data for training algorithms (Hassan et al., 2012). During next few decades, large-scale data management will become collaboration between machines and humans.

The simplest form of automation consists of scripting curation activities that are recurrent, creating specialized curation agents. This approach is used, for example, in Wikipedia (Wiki Bots) for article cleaning and detecting vandalism. Another automation process consists in providing an algorithmic approach for the validation or annotation of the data against reference standards (Data Curation Interview: Antony Williams, 2013). This would contribute to a *"likesonomy"* where both humans and algorithms could provide further evidence in favour or against data (Data Curation Interview: Antony Williams, 2013). These approaches provide a way to automate more recurrent parts of the curation tasks and can be implemented today in any curation pipeline (there are no major technological barriers). However, the construction of these algorithmic or reference bases has a high cost effort (in terms of time consumption and

[1] http://openrefine.org/

expertise), since they depend on an explicit formalization of the algorithm or the reference criteria (rules).

Data Curation Automation: More sophisticated automation approaches that could alleviate the need for the explicit formalization of curation activities will play a fundamental role in reducing the cost of data curation. The research areas that can mostly impact data curation automation are:

- **Curating by Demonstration(CbD)/Induction of Data Curation Workflows:** Programming by example (or programming by demonstration (PbD)) (Cypher, 1993; Flener, 2008; Lieberman, 2001) is a set of end-user development approaches in which the user actions on concrete instances are generalized into a program. PbD can be used to allow distribution and amplification of the system development tasks by allowing users to become programmers. Despite being a traditional research area, and on the research on PbD data curation platforms (Tuchinda et al., 2007; Tuchinda, 2011), PbD methods have not been extensively applied into data curation systems.
- **Evidence-based Measurement Models of Uncertainty over Data:** The quantification and estimation of generic and domain specific models of uncertainty from distributed and heterogeneous evidence bases can provide the basis for the decision on what should be delegated or validated by humans and what can be delegated to algorithmic approaches. IBM Watson is an example of system that uses at its centre a statistical model to determine the probability of an answer of being correct (Ferruci et al., 2008). Uncertainty models can also be used to route tasks according to level of expertise, minimizing the cost and maximizing the quality of data curation.

Both areas have a strong connection with the application of machine learning in the data curation field.

Curation at source: *Sheer curation or curation-at-source*, is an approach to curate data where lightweight curation activities are integrated into the normal workflow of those creating and managing data and other digital assets. (Curry et al., 2010). Sheer curation activities can include lightweight categorisation and normalisation activities. An example would be, vetting or "rating" the results of a categorization process performed by a curation algorithm. Sheer curation activities can also be composed with other curation activities, allowing more immediate access to curated data while also ensuring the quality control that is only possible with an expert curation team.

The following are the high-level objectives of sheer curation described by (Hedges & Blanke, 2012):

- Avoid data deposit by integrating with normal workflow tools
- Capture provenance information of the workflow
- Seamless interfacing with data curation infrastructure

State-of-the-Art Data Curation Platforms

- **Data Tamer[1]:** This prototype aims to replace the current developer-centric extract-transform-load (ETL) process with automated data integration. The system uses a suit of algorithms to automatically map schemas and de-duplicate entities. However, human experts and crowds are leveraged to verify integration updates that are particularly difficult for algorithms.
- **ZenCrowd[1]:** This system tries to address the problem of linking named entities in text with a knowledge base. ZenCrowd bridges the gap between automated and manual linking by improving the results of automated linking with humans. The prototype was demonstrated for linking named entities in news articles with entities in Linked Open Data cloud.
- **CrowdDB[1]:** This database system answers SQL queries with the help of crowds. Specifically, queries that cannot be answered by a database management system or a search engine. As opposed to the exact operation in databases, CrowdDB allows fuzzy operations with the help of humans. For example, ranking items by relevance or comparing equivalence of images.
- **Qurk[1]:** Although similar to CrowdDB, this system tries to improve costs and latency of human-powered sorts and joins. In this regard, Qurk applies techniques such as batching, filtering, and output agreement.
- **Wikipedia Bots:** Wikipedia runs scheduled algorithms to access quality of text articles, known as Bots. These bots also flag articles that require further review by experts. SuggestBot[1] recommends flagged articles to a Wikipedia editor based on their profile.

7.4. Human-Data Interaction

Focus on the interactivity, easy of curation actions: Data interaction approaches which facilitate data transformation and access are fundamental for expanding the spectrum of data curators' profiles. *There are still major barriers for interacting with structured data* and the process of querying, analysing and modifying data inside databases is in most cases mediated by IT professionals or domain-specific applications. Supporting domain experts and casual users in querying, navigating, analysing and transforming structured data is a fundamental functionality in data curation platforms.

According to Carole Goble "*from a Big Data perspective, the challenges are around finding the slices, views or ways into the dataset that enables you to find the bits that need to be edited, changed*" (Data Curation Interview: Carole Goble, 2013). Therefore, appropriate summarization and visualization of data is important not only from the usage perspective but also from maintenance perspective (Hey & Trefethen, 2004). Specifically, for the collaborative methods of data cleaning, it is fundamental to enable discovery of anomalies in both structured and unstructured data. Additionally, making data management activities more mobile and interactive is required as mobile devices overtake desktops. The following technologies provide direction towards better interaction:

- **Data-Driven Documents[1] (D3.js):** D3.js is library for displaying interactive graphs in web documents. This library adheres to open web standard such as HTML5, SVG and CSS, to enable powerful visualizations with open source licensing.
- **Tableau[2]:** This software allows users to visualize multiple dimensions of relational databases. Furthermore it enables visualization of unstructured data through third-party adapters. Tableau has received a lot of attention due to its ease of use and free access public plan.
- **Open Refine[3]:** This open source application allows users to clean and transform data from variety of formats such as CSV, XML, RDF, JSON, etc. Open Refine is particularly useful for finding outliers in data and checking distribution of values in columns through facets. It allows data reconciliation with external data sources such as Freebase and OpenCorporates[4].

Structured query languages such as SQL are the default approach for interacting with databases, together with graphical user interfaces which are developed as a façade over structured query languages. The query language syntax and the need to understand the schema of the database are not appropriate for domain experts to interact and explore the data. Querying progressively more complex structured databases and dataspaces will demand different approaches suitable for different tasks and different levels of expertise (Franklin et al., 2005). New approaches for interacting with structured data have evolved from the early research stage and can provide the basis for new suites of tools which can facilitate the interaction between user and data. Examples are keyword search, visual query interfaces and natural language query interfaces over databases (Franklin et al., 2005; Freitas et al. 2012; Kaufman & Bernstein, 2007). Flexible approaches for database querying depends on the ability of the approach to interpret the user query intent, matching it with the elements in the database. These approaches are ultimately dependent on the creation of semantic models that support semantic approximation (Freitas et al. 2011). Despite going beyond the proof-of-concept stage these functionalities and approaches have not migrated to commercial-level applications.

1 http://d3js.org/

2 http://www.tableausoftware.com/public/

3 https://github.com/OpenRefine/OpenRefine/wiki

4 https://www.opencorporates.com

7.5. Trust

Capture of data curation decisions & provenance management: As data reuse grows, the consumer of third-party data needs to have mechanisms in place to verify the trustworthiness and the quality of the data. Some of the data quality attributes can be evident by the data itself, while others depend on an understanding of the broader context behind the data, i.e. the provenance of the data, the processes, artefacts and actors behind the data creation.

Capturing and representing the context in which the data was generated and transformed and making it available for data consumers is a major requirement for data curation for datasets targeted towards third-party consumers. Provenance standards such as W3C PROV[1] provide the grounding for the interoperable representation of the data. However, data curation applications still need to be instrumented to capture provenance. Provenance can be used to explicitly capture and represent the curation decisions that are made (Data Curation Interview: Paul Groth, 2013). However, there is still a relatively low adoption on provenance capture and management in data applications. Additionally, manually evaluating trust and quality from provenance data can be a time consuming process. The representation of provenance needs to be complemented by automated approaches to derive trust and assess data quality from provenance metadata, under the context of a specific application.

Fine-grained permission management models and tools: Allowing large user bases to collaborate demands the creation of fine-grained permission/rights associated with curation roles. Most systems today have a coarse-grained permission system, where system stewards oversee general contributors. While this mechanism can fully address the requirements of some projects, there is a clear demand for more fine-grained permission systems, where permissions can be defined at a data item level (Qin & Atluri, 2003; Ryutov et al., 2009) and can be assigned in a distributed way. In order to support this fine-grained control, the investigation and development of automated methods for permissions inference and propagation (Kirrane et al., 2013), as well as low-effort distributed permission assignment mechanisms, is of primary importance. Analogously, similar methods can be applied to a fine-grained control of digital rights (Rodrıguez-Doncel et al., 2013).

7.6. Standardization & Interoperability

Standardized data model and vocabularies for data reuse: A large part of the data curation effort consists of integrating and repurposing data created under different contexts. In many cases this integration can involve hundreds of data sources. Data model standards such as the Resource Description Framework (RDF)[2] facilitate the data integration at the data model level. The use of Universal Resource Identifiers (URIs) in the identification of data entities works as a Web-scale open foreign key, which promotes the reuse of identifiers across different datasets, facilitating a distributed data integration process.

The creation of terminologies and vocabularies is a critical methodological step in a data curation project. Projects such as the NYT Index (Curry et al., 2010) or the ProteinDataBank (Bernstein, 1977) prioritize the creation and evolution of a vocabulary that can serve to represent and annotate the data domain. In the case of PDB, the vocabulary expresses the representation needs of a community. The use of shared vocabularies is part of the vision of the Linked Data Web (Berners-Lee, 2009) and it is one methodological tool which can be used to facilitate semantic interoperability. While the creation of a vocabulary is more related to a methodological dimension, semantic search, schema mapping or ontology alignment approaches (Shvaiko & Euzenat, 2005; Freitas et al. 2012) are central for reducing the burden of manual vocabulary mapping on the end user side, reducing the burden for terminological reuse (Freitas et al., 2012).

[1] http://www.w3.org/TR/prov-primer/
[2] http://www.w3.org/TR/rdf11-primer/

Better integration and communication between curation tools: Data is created and curated in different contexts and using different tools (which are specialised to satisfy different data curation needs). For example a user may analyse possible data inconsistencies with a visualization tool, do schema mapping with a different tool and then correct the data using a crowdsourcing platform. The ability to move the data seamlessly between different tools and capture user curation decisions and data transformations across different platforms is fundamental to support more sophisticated data curation operations that may demand highly specialised tools and to make the final result trustworthy (Data Curation Interview: Paul Groth, 2013; Data Curation Interview: James Cheney, 2013). The creation of standardised data models and vocabularies (such as W3C PROV) addresses part of the problem. However, data curation applications need to be adapted to capture and manage provenance and to provide better adoption over existing standards.

7.7. Data Curation Models

Minimum information models for data curation: Despite recent efforts in the recognition and understanding behind the field of data curation (Palmer et al., 2013; Lord et al., 2004), the processes behind it are still to be better formalized. The adoption of methods such as minimum information models (La Novere et al., 2008) and their materialization in tools is one example of methodological improvement that can provide a minimum quality standard for data curators. In eScience, MIRIAM (Minimum Information Required In The Annotation of Models) (Laibe & Le Novère, 2007) is an example of a community-level effort to standardize the annotation and curation processes of quantitative models of biological systems.

Curating Nanopublications: coping with the long tail of science: With the increase in the amount of scholarly communication, it is increasingly difficult to find, connect and curate scientific statements (Mons et al., 2009; Groth et al., 2010). Nanopublications are core scientific statements with associated contexts (Groth et al., 2010), which aims at providing a synthetic mechanism for scientific communication. Nanopublications are still an emerging paradigm, which may provide a way for the distributed creation of semi-structured data in both scientific and non-scientific domains.

Investigation of theoretical principles and domain specific models for data curation: Models for data curation should evolve from the ground practice into a more abstract description. The advancement of automated data curation algorithms will depend on the definition of theoretical models and on the investigation of the principles behind data curation (Buneman et al., 2008). Understanding the causal mechanisms behind workflows (Cheney, 2010) and the generalization conditions behind data transportability (Pearl & Bareinboim, 2011) are examples of theoretical models which can impact data curation, guiding users towards the generation and representation of data which can be reused in broader contexts.

7.8. Unstructured & Structured Data Integration

Entity recognition and linking: Most of the information on the Web and in organizations is available as unstructured data (text, videos, etc.). The process of making sense of information available as unstructured data is time consuming: differently from structured data, unstructured data cannot be directly compared, aggregated and operated. At the same time, unstructured data holds most of the information of the *long tail of data variety* (Figure 3-1).

Extracting structured information from unstructured data is a fundamental step for making the long tail of data analysable and interpretable. Part of the problem can be addressed by information extraction approaches (e.g. relation extraction, entity recognition and ontology extraction) (Freitas et al., 2012; Schutz & Buitelaar, 2005; Han et al. 2011; Data Curation Interview: Helen Lippell, 2013). These tools extract information from text and can be used to

automatically build semi-structured knowledge from text. There are information extraction frameworks that are mature to certain classes of information extraction problems, but their adoption remains limited to early-adopters (Curry et al. 2010; Data Curation Interview: Helen Lippell, 2013).

Use of open data to integrate structured & unstructured data: Another recent shift in this area is the availability of large-scale structured data resources, in particular open data, which is supporting information extraction. For example entities in open datasets such as DBpedia (Auer et al., 2007) and Freebase (Bollacker et al. 2008) can be used to identify named entities (people, places and organizations) in texts, which can be used to categorize and organize text contents. Open data in this scenario works as a common sense knowledge base for entities and can be extended with domain specific entities inside organisational environments. Named entity recognition and linking tools such as DBpedia Spotlight (Mendes et al., 2011) can be used to link structured and unstructured data.

Complementarily, unstructured data can be used to provide a more comprehensive description for structured data, improving content accessibility and semantics. *Distributional semantic models*, semantic models which are built from large-scale collections (Freitas et al. 2012), can be applied to structured databases (Freitas & Curry, 2014) and are examples of approaches which can be used to enrich the semantics of the data.

Natural language processing pipelines: The Natural Language Processing (NLP) community has matured approaches and tools that can be directly applied to projects which demand dealing with unstructured data. Open Source projects such as Apache UIMA[1] facilitates the integration of NLP functionalities into other systems. Additionally, strong industry use cases such as IBM Watson (Ferrucci et al., 2013), Thomson Reuters, The New York Times (Curry et al., 2013), Press Association (Data Curation Interview: Hellen Lippell) are shifting the perception of NLP techniques from the academic to the industrial field.

Requirement Category	Emerging Approach	Adoption/Status	Exemplar Use Case
Incentives Creation & Social Engagement Mechanisms	Open and interoperable data policies	Early-stage / Limited adoption	Data.gov.uk
	Better recognition of the data curation role	Lacking of adoption / Despite of the exemplar use cases, the data curator role is still not recognised	Chemspider, Wikipedia, ProteinDataBank
	Attribution and recognition of data and infrastructure contributions	Standards emerging / Adoption missing	Altmetrics (Priem et al., 2010), ORCID
	Better understanding of social engagement mechanisms	Early-stage	GalaxyZoo (Forston et al., 2011), Foldit (Khatib et al., 2011)
	Pre-competitive partnerships	Seminal use cases	Pistoia Alliance (Wise, 2012)

[1] http://uima.apache.org/

Economic Models	Public-private partnerships	Seminal use cases	Geoconnections (Harper, 2012)
	Quantification of the economic impact of data	Seminal use cases	JISC, 2011 (“Data centres: their use, value and impact”)
Curation at Scale	Human computation & Crowdsourcing services	Industry-level adoption / Services are available but there is space for market specialization	CrowdFlower, Amazon Mechanical Turk
	Evidence-based Measurement Models of Uncertainty over Data	Research stage	IBM Watson (Ferrucci et al. 2010)
	Programming by demonstration, induction of data transformation workflows	Research stage / Fundamental research areas are developed. Lack of applied research in a workflow & data curation context.	Tuchinda et al., 2007; Tuchinda, 2011
	Curation at source	Existing use cases both in academic projects and industry	The New York Times
	General-purpose data curation pipelines	Available Infrastructure	OpenRefine, Karma, Scientific Workflow management systems
	Algorithmic validation/annotation	Early stage	Wikipedia, Chemspider
Human-Data Interaction	Focus on the interactivity, easy of actions	Seminal tools available	OpenRefine
	Natural language interfaces, Schema-agnostic queries	Research stage	IBM Watson (Ferrucit et al., 2010), Treo (Freitas & Curry, 2014)
Trust	Capture of data curation decisions	Standards are in place, instrumentation of applications needed	OpenPhacts
	Fine-grained permission management models and tools	Coarse-grained infrastructure available.	Qin & Atluri, 2003; Ryutov et al., 2009; Kirrane et al., 2013; Rodrıguez-Doncel et al., 2013
	Standardized data model	Standards are available.	RDF(S), OWL
	Reuse of vocabularies	Technologies for supporting	

Standardization & Interoperability		vocabulary reuse is needed	Linked Open Data Web (Berners-Lee, 2009)
	Better integration and communication between tools	Low	N/A
	Interoperable provenance representation	Standard in place / Standard adoption is still missing	W3C PROV
Curation Models	Definition of minimum information models for data curation	Low adoption	MIRIAM (Laibe & Le Novère, 2007)
	Nanopublications	Emerging concept	Mons et al .2009, Groth et al. 2010
	Investigation of theoretical principles and domain specific models for data curation	Emerging concept	Pearl & Bareinboim, 2011
Unstructured-Structured Integration	NLP Pipelines	Tools are available, Adoption is low	IBM Watson (Ferrucci et al., 2010)
	Entity recognition and alignment	Tools are available, Adoption is low	DBpedia Spotlight (Mendes et al., 2011), IBM Watson (Ferrucci et al., 2010)

Table 7-1 Emerging approaches for addressing the future requirements.

8. Sectors Case Studies for Big Data Curation

In this section, we discuss case studies that cover different data curation processes over different domains. The purpose behind the case studies is to capture the different workflows that have been adopted or designed in order to deal with data curation in the Big Data context.

8.1. Health and Life Sciences

ChemSpider: ChemSpider[1] is a search engine that provides free service access to the structure-centric chemical community. It has been designed to aggregate and index chemical structures and their associated information into a single searchable repository. ChemSpider contains tens of millions of chemical compounds and its associated data, and is serving as a data provider to websites and software tools. Available since 2007, ChemSpider has collated over 300 data sources from chemical vendors, government databases, private laboratories and individuals, providing access to millions of records related to chemicals. Used by chemists for identifier conversion and properties predictions, ChemSpider datasets are also heavily leveraged by chemical vendors and pharmaceutical companies as pre-competitive resources for experimental and clinical trial investigation.

Data curation in ChemSpider consists in the manual annotation and correction of data (Pence et al., 2010). This may include changes to the chemical structures of a compound, addition or deletion of identifiers associated with a chemical compound, associating links between a chemical compound and its related data sources etc. ChemSpider supports two different ways for curators to help in curating data at ChemSpider:

- Post comments on a record in order to take appropriate action on your concern by the Master curator.
- As a registered member with curation rights, allows to participate directly in marking data for master curation or to remove erroneous data.

ChemSpider adopts a meritocratic 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 period of contribution. The platform has a blended human and computer-based curation process. Robotic Curation uses algorithms for error correction and data validation at deposition time.

ChemSpider uses a mixture of computational approaches to perform certain level of data validation. They have built their own chemical data validation tool, which is called CVSP (Chemical Validation and Standardization Platform). CVSP helps chemists to check chemicals and tell quickly whether or not they are validly represented, whether there are any data quality issues so that they can flag those quality issues easily and efficiently.

Using the Open Community model, ChemSpider distributes its curation activity across its community using crowdsourcing to accommodate massive growth rates and quality issues. They use a wiki-like approach for people to interact with the data, so that they can annotate it, validate it, curate it, delete it, flag it if deleted. ChemSpider is also in the phase of implementing an automated recognition system that will measure the contribution effort of curators through the data validation and engagement process. The contribution metrics becomes then publicly viewable and accessible through a central RSC profile as shown in Figure 8-1 RSC profile of a curator with awards attributed based on his/her contributions.Figure 8-1.

1 http://www.chemspider.com

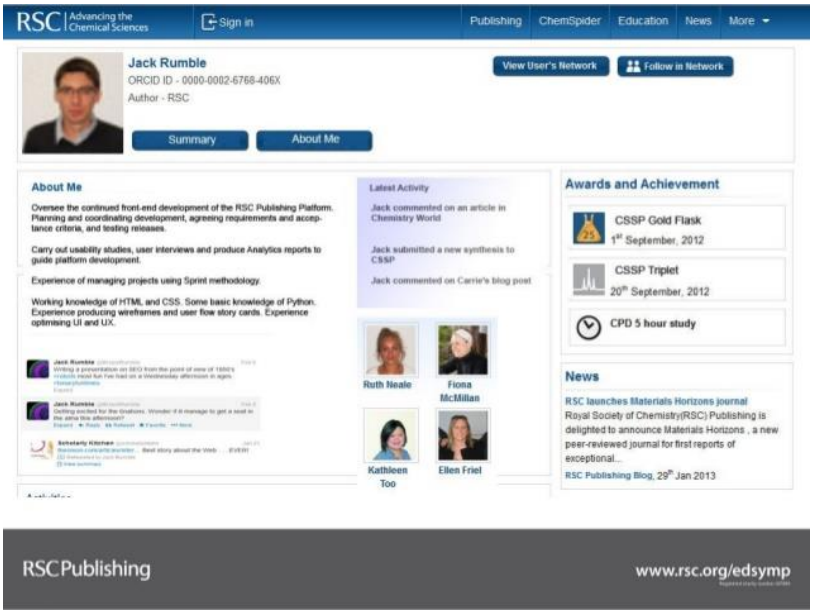

Figure 8-1 RSC profile of a curator with awards attributed based on his/her contributions.

Protein Data Bank: The Research Collaboratory for Structural Bioinformatics Protein Data Bank[1] (RCSB PDB) is a group dedicated to improve understanding of the functions of biological systems through the study of 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. The PDB has had over 300 million data set downloads. Its tools and resource offerings have grown from a curated data download service, to a platform that serves molecular visualization, search, and analysis tools.

A significant amount of the curation process at PDB consists in providing standardised vocabulary for describing the relationships between biological entities, varying from organ tissue to the description of the molecular structure. 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.

The data curation process also covers the identification and correction of inconsistencies over the 3-D protein structure and experimental data. The platform accepts the deposition of data in multiple formats such as the legacy PDB format, mmCif, and the current PDBML. In order to implement a global hierarchical governance approach to the data curation workflow, wwPDB staff review and annotate each submitted entry before robotic curation checks for plausibility as part of the data deposition, processing and distribution. The data curation effort is distributed across their sister sites.

Robotic curation automates the data validation and verification. Human curators contribute to the definition of rules for the detection of inconsistencies. The curation process is also propagated retrospectively, where errors found in the data are corrected retrospectively to the archives. Up to date versions of the data sets are released on weekly basis to keep all sources consistent with the current standards and to ensure good data curation quality.

FoldIt: Foldit (Good & Su, 2009) is a popular example of human computation applied to a complex problem, i.e. finding patterns of protein folding. The developers of Foldit have used gamification to enable human computation. Through these games people can predict protein structure which might help in targeting drugs at particular disease. Current computer algorithms are unable to deal with the exponentially high number of possible protein structures. To

1 http://www.pdb.org

overcome this problem, Foldit uses competitive protein folding to generate best proteins (Eiben et al., 2012) (Figure 8-2).

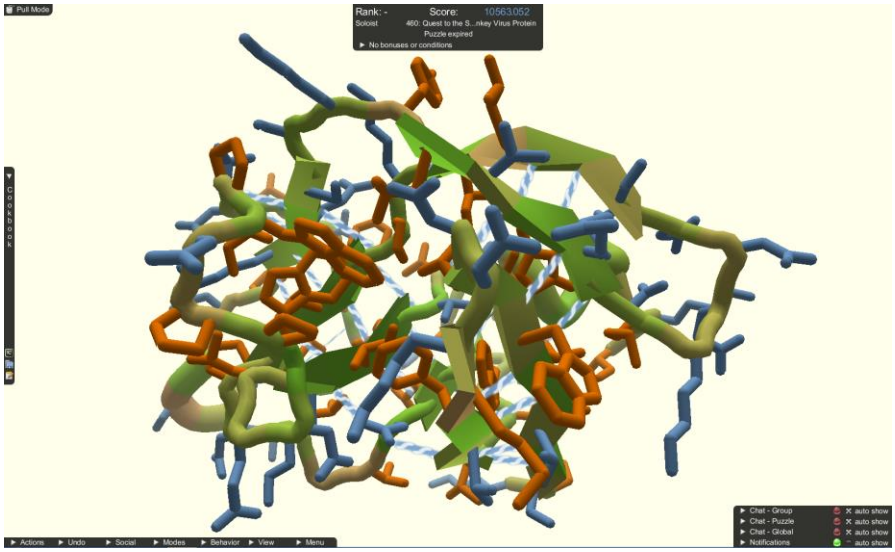

Figure 8-2 An example solution to a protein folding problem with Fold.it[1]

8.2. Telco, Media, Entertainment

Press Association: Press Association (PA) is the national news agency for the UK and Ireland and a leading multimedia content provider across Web, mobile, broadcast and print. For the last 145 years, PA has been providing feeds of text, data, photos and videos, to all major UK media outlets as well as corporate customers and the public sector.

The objective of data curation at Press Association 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 Press Association employs a large number of curators in the content classification process, working over a large number of data sources. A curator inside Press Association is an analyst, who collects, aggregates, classifies, normalizes, and analyses the raw information coming from different data sources. Since the nature of the information analysed at Press Association 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. In the curation process, automatic tools provide a first level triage and classification, which is further refined by the intervention of human curators as shown in Figure 8-3.

The data curation process starts with an article submitted to a platform which uses a set of linguistic extraction rules over unstructured text to automatically derive tags for the article, enriching it with machine readable structured data. A data curator then selects the terms that better describe the contents and inserts new tags if necessary. The tags enrich the original text with the general category of the analysed contents, while also providing a description of specific entities (places, people, events, facts) that are present in the text. The meta-data manager then reviews the classification and the content is published online.

Thomson Reuters: Thomson Reuters is a leading information provider company which is focused on the provision of specialist curated information in different domains, including Healthcare, Science, Financial, Legal and Media.

In addition to the selection and classification of the most relevant information for its customers, Thomson Reuters focuses on the deployment of information (including structured data) in a way

[1] Image courtesy www.fold.it

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 selection and classification, which is further refined by the intervention of human curators. A typical curator is a domain specialist, who selects, aggregates, classifies, normalises and analyses the raw information coming from different data sources. Semantic Web technologies are already applied in the company's data environment. The overall data curation workflow at Thomson Reuters is depicted in Figure 8-4.

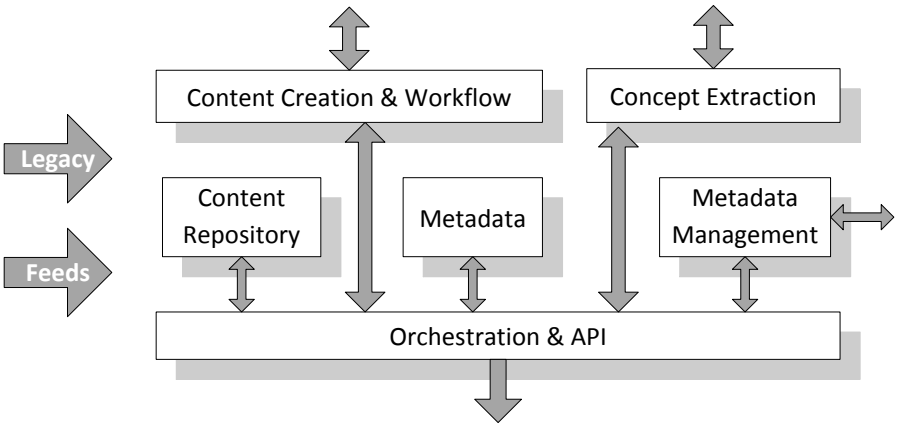

Figure 8-3: PA Content and Metadata Pattern Workflow.

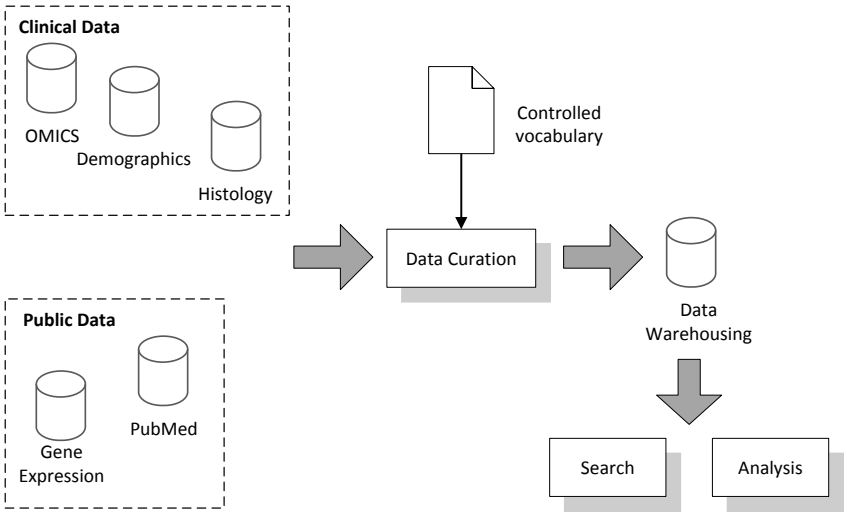

Figure 8-4: A typical data curation process at Thomson Reuters.

The New York Times: The New York Times (NYT) is the largest metropolitan and the third largest newspaper in the United States. The company has a long history of the curation of its articles in its 100-year-old curated repository (NYT Index).

New York Times curation pipeline (see Figure 8-5) starts with an article getting out of the newsroom. The first level curation consists in the content classification process done by the editorial staff, which consists of several hundred journalists. 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[1]). 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

1 SAS Teragram http://www.teragram.com

Index Department. Teragram suggests tags based on the Index vocabulary that can potentially describe the content of the article (Curry et al, 2010). The member of the editorial staff then selects the terms that better describe the contents and inserts new tags if necessary.

Taxonomy managers review the classification and the content is published online, providing continuous feedback into the classification process. 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. The data curation workflow at NYT is outlined in Figure 8-5.

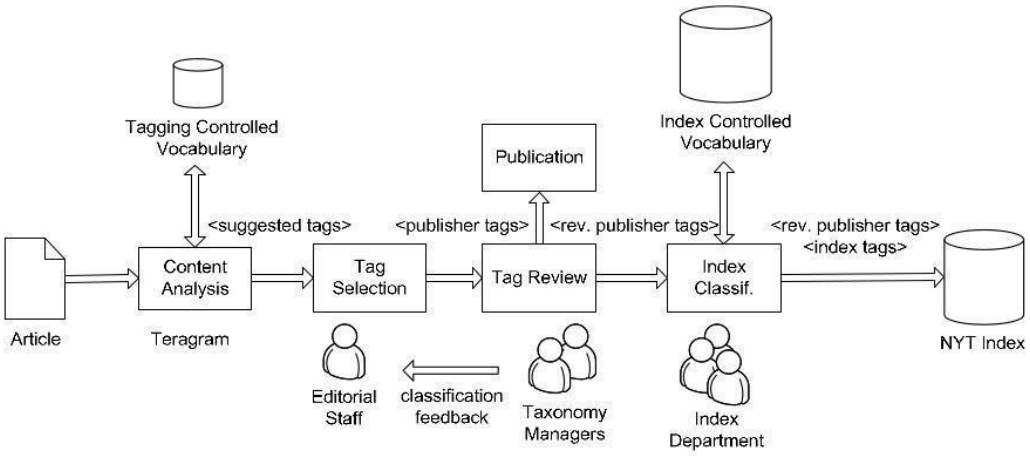

Figure 8-5: The NYT article classification curation workflow.

8.3. Retail

Ebay: Ebay is one of the most popular online marketplaces that caters for millions of products and customers. Ebay has employed human computation to solve two important issues of data quality; managing product taxonomies and finding identifiers in product descriptions. Crowdsourced workers helped Ebay in improving the speed and quality of product classification algorithms at lower costs (Lu et al., 2010) (Figure 8-6).

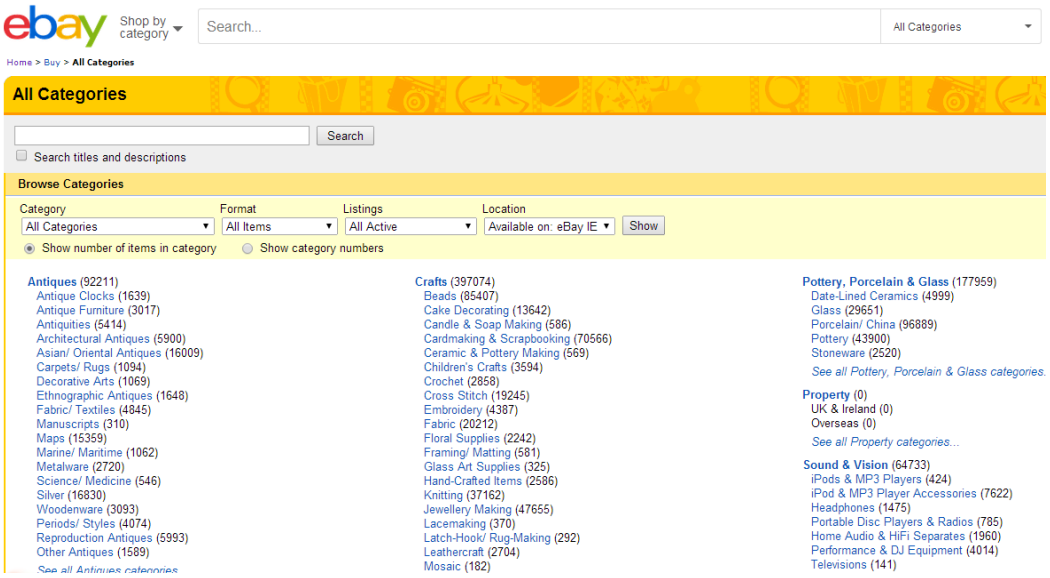

Figure 8-6 Taxonomy of products used by Ebay to categorize items with help of crowdsourcing.

Unilever: Unilever is one of the world's largest manufacturers of consumer goods, with global operations. Unilever utilized crowdsourced human computation for two problems related to their marketing strategy for new products. Human computation was used to gather sufficient data about customer feedback and to analyze public sentiment of social media. Initially Unilever developed a set of machine learning algorithms to do analysis sentiment of customers across their product range. However these sentiment analysis algorithms were unable to account for regional and cultural differences between target populations. Therefore, Unilver effectively improved the accuracy of sentiment analysis algorithms with crowdsourcing, by verifying the output algorithms and gathering feedback from an online crowdsourcing platform, i.e. Crowdflower.

9. Conclusions

With the growth in the number of data sources and of decentralised content generation, ensuring data quality becomes a fundamental issue on data management environments in the Big Data era. The evolution of data curation methods and tools is a cornerstone element for ensuring data quality at the scale of Big Data.

Based on the evidence collected by an extensive survey that included a comprehensive literature analysis, interviews with data curation experts, questionnaires and case studies, this whitepaper aimed at depicting the future requirements and emerging trends for data curation. This analysis can provide to data curators, technical managers and researchers an up-to-date view of the challenges, approaches and opportunities for data curation in the Big Data era.

10. Acknowledgements

The authors would like to thank Nur Aini Rakhmawati and Aftab Iqbal (NUIG) for their contribution to the first versions of the whitepaper and interview transcriptions and Helen Lippell (PA) for her review and feddback.

11. References

Alonso, O., Baeza-Yates, R. (2011) Design and implementation of relevance assessments using crowdsourcing. *Advances in information retrieval*, 153-164.

Armstrong, A. W. et al. "Crowdsourcing for research data collection in rosacea." *Dermatology online journal* 18.3 (2012).

Aroyo, Lora, and Chris Welty. "Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard." *WebSci2013. ACM* (2013).

Auer et al., DBpedia: a nucleus for a web of open data. In Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference, 722-735, (2007).

Ball, A, *Preservation and Curation in Institutional Repositories.* Digital Curation Centre, (2010).

Beagrie, N., Chruszcz, J., Lavoie, B., *Keeping Research Data Safe: A cost model and guidance for UK universities.* JISC, (2008).

Berners-Lee, T., Linked Data Design Issues, http://www.w3.org/DesignIssues/LinkedData.html, (2009).

Bernstein et al., The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures , J Mol Biol.112(3):535-42 (1977).

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S., DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web, 7*(3), 154-165, (2009).

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J., Freebase: a collaboratively created graph database for structuring human knowledge. *Proceedings of the ACM SIGMOD International Conference on Management of Data,* pp. 1247-1250. New York, NY, (*2008)*.

Brodie, M. L., Liu, J. T. The power and limits of relational technology in the age of information ecosystems. On The Move Federated Conferences, (2010).

Buneman, P., Chapman, A., & Cheney, J., Provenance management in curated databases, *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data* (pp. 539-550), (2006).

Buneman, P., Cheney, J., Tan, W., Vansummeren, S., Curated Databases, in Proceedings of the Twenty-seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, (2008).

Cheney, J., Causality and the semantics of provenance, (2010).

Curry, E., Freitas, A., & O'Riáin, S., The Role of Community-Driven Data Curation for Enterprise. In D. Wood, *Linking Enterprise Data* pp. 25-47, (2010).

Cragin, M., Heidorn, P., Palmer, C. L.; Smith, Linda C., An Educational Program on Data Curation, ALA Science & Technology Section Conference, (2007).

Crowdsourcing: utilizing the cloud-based workforce, Whitepaper, (2012)

Cypher, A., Watch What I Do: Programming by Demonstration, (1993).

Data centres: their use, value and impact, JISC Report, (2011).

Doan, A., Ramakrishnan, R., Halevy, A.. Crowdsourcing systems on the world-wide web. *Communications of the ACM* 54.4, 86-96, (2011).

Eastwood, T., Appraising digital records for long-term preservation. *Data Science Journal, 3*, 202-208, (2004).

Eiben, C. B. et al. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nature biotechnology,* 190-192, (2012).

European Journal of Biochemistry, Volume 80, Issue 2, pp. 319–324, (1977).

Forston et al., Galaxy Zoo: Morphological Classification and Citizen Science, Machine learning and Mining for Astronomy, (2011).

Freitas, A., Curry, E., Oliveira, J. G., O'Riain, S., Querying Heterogeneous Datasets on the Linked Data Web: Challenges, Approaches and Trends. *IEEE Internet Computing, 16*(1), 24-33, (2012).

Ferrucci et al., Building Watson: An Overview of the DeepQA Project, AI Magazine, (2010).

Finin, Tim et al., Annotating named entities in Twitter data with crowdsourcing, *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*: 80-88, (2010).

Flener, P., Schmid, U., An introduction to inductive programming, Artif Intell Rev 29:45–62, (2008).

Franklin, M., Halevy, A., Maier, D., From databases to dataspaces: a new abstraction for information management, ACM SIGMOD Record, Volume 34, Issue 4, pp. 27-33, (2005).

Freitas, A., Oliveira, J.G., O'Riain, S., Curry, E., Pereira da Silva, J.C, Querying Linked Data using Semantic Relatedness: A Vocabulary Independent Approach, In Proceedings of the 16th

International Conference on Applications of Natural Language to Information Systems (NLDB), (2011).
Freitas, A., Carvalho, D., Pereira da Silva, J.C., O'Riain, S., Curry, E., A Semantic Best-Effort Approach for Extracting Structured Discourse Graphs from Wikipedia, In Proceedings of the 1st Workshop on the Web of Linked Entities (WoLE 2012) at the 11th International Semantic Web Conference (ISWC), (2012).
Freitas, A., Curry, E., Natural Language Queries over Heterogeneous Linked Data Graphs: A Distributional-Compositional Semantics Approach, In Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI), Haifa, (2014).
Gartner, 'Dirty Data' is a Business Problem, Not an IT Problem, says Gartner, Press release, (2007).
Gil et al., Mind Your Metadata: Exploiting Semantics for Configuration, Adaptation, and Provenance in Scientific Workflows, In Proceedings of the 10th International Semantic Web Conference (ISWC), (2011).
Giles, J., Learn a language, translate the web, *New Scientist,* 18-19, (2012).
Groth, P., Gibson, A., Velterop, J., The anatomy of a nanopublication. Inf. Serv. Use 30, 1-2, 51-56, (2010).
Harper, D., GeoConnections and the Canadian Geospatial Data Infrastructure (CGDI): An SDI Success Story, Global Geospatial Conference, (2012).
Hedges, M., & Blanke, T., Sheer curation for experimental data and provenance. *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pp. 405–406, (2012).
Higgins, S. The DCC Curation Lifecycle Model. *The International Journal of Digital Curation, 3*(1), 134-140, (2008).
Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Yon Rhee, S., Big data: The future of biocuration. *Nature, 455*(7209), 47-50, (2008).
Jakob, M., García-Silva, A., Bizer C., DBpedia spotlight: shedding light on the web of documents, Proceedings of the 7th International Conference on Semantic Systems, Pages 1-8, 2011.
Kaggle: Go from Big Data to Big Analytics, <http://www.kaggle.com/>, (2005).
Kaufmann, E., Bernstein, A., How Useful are Natural Language Interfaces to the Semantic Web for Casual End-users?, Proceedings of the 6th international The semantic web conference, 2007, p. 281-294.
Khatib et al., Crystal structure of a monomeric retroviral protease solved by protein folding game players, Nature Structural & Molecular Biology 18, 1175–1177, (2011).
Kirrane, S., Abdelrahman, A., Mileo, S., Decker, S., Secure Manipulation of Linked Data, In Proceedings of the 12th International Semantic Web Conference, (2013).
Kittur, Aniket et al. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web* 1.2 (2007).
Knight, S.A., Burn, J., Developing a Framework for Assessing Information Quality on the World Wide Web. Informing Science. 8: pp. 159-172, 2005.
Kong, N., Hanrahan, B., Weksteen, T., Convertino, G., Chi, E. H., VisualWikiCurator: Human and Machine Intelligence for Organizing Wiki Content. *Proceedings of the 16th International Conference on Intelligent User Interfaces*, pp. 367-370, (2011).
La Novere et al., Minimum information requested in the annotation of biochemical models (MIRIAM), Nat Biotechnol , 23(12), 1509-15, (2005).
Law, E., von Ahn, L., Human computation, *Synthesis Lectures on Artificial Intelligence and Machine Learning*: 1-121, (2011).
Laibe, C., Le Novère, N., MIRIAM Resources: Tools to generate and resolve robust cross-references in Systems Biology, BMC Systems Biology 1: 58, (2007).
Law, E., von Ahn, L., Human computation, *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 1-121, (2011).
Law, E., von Ahn, L., Input-agreement: a new mechanism for collecting data using human computation games. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 4, 1197-1206, (2009).
Lieberman, H., Your Wish is My Command: Programming By Example, (2001).
Lord, P., Macdonald, A., *e-Science Curation Report.* JISC, (2003).
Lord, P., Macdonald, A., Lyon, L., Giaretta, D., From Data Deluge to Data Curation, (2004).
Mons, B., Velterop, J., Nano-Publication in the e-science era, International Semantic Web Conference, (2009).
Morris, H.D., Vesset, D., Managing Master Data for Business Performance Management: The Issues and Hyperion's Solution, Technical Report, (2005).
Norris, R. P., How to Make the Dream Come True: The Astronomers' Data Manifesto, (2007).

Palmer et al., Foundations of Data Curation: The Pedagogy and Practice of “Purposeful Work” with Research Data, (2013).
Pearl, J., Bareinboim, E., Transportability of causal and statistical relations: A formal approach, in Proceedings of the 25th National Conference on Artificial Intelligence (AAAI), (2011).
Pence, H. E., & Williams, A., ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education, 87*(11), 1123-1124, (2010).
Priem, J., Taraborelli, D., Groth, P. Neylon, C., Altmetrics: A manifesto, http://altmetrics.org/manifesto/, (2010).
Han, X., Sun, L., Zhao, J., Collective Entity Linking in Web Text: A Graph-based Method, Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, (2011).
Verhaar, P., Mitova, M., Rutten, P., Weel, A. v., Birnie, F., Wagenaar, A., Gloerich, J., *Data Curation in Arts and Media Research.* SURFfoundation, (2010).
von Ahn, L., and Laura Dabbish. Designing games with a purpose. *Communications of the ACM* 51.8, 58-67, (2008).
von Ahn, L., Duolingo: learn a language for free while helping to translate the web, *Proceedings of the 2013 international conference on Intelligent user interfaces* 19, 1-2, (2013).
Ipeirotis, P. G., Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students* 17.2, 16-21, (2010).
Sheth, A., Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics, Interoperating Geographic Information Systems The Springer International Series in Engineering and Computer Science Volume 495, pp 5-29, (1999).
Stonebracker et al. Data Curation at Scale: The Data Tamer System, 6th Biennial Conference on Innovative Data Systems Research (CIDR), (2013).
Surowiecki, James. *The wisdom of crowds*. Random House LLC, (2005).
Wang, R., Strong, D., Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4): p. 5-33, (1996).
Qin, L., Atluri, V., Concept-level access control for the Semantic Web. In Proceedings of the ACM workshop on XML security - XMLSEC '03. ACM Press, (2003).
Rodrıguez-Doncel, V., Gomez-Perez, A., Mihindukulasooriya, N., Rights declaration in Linked Data, in Proceedings of the Fourth International Workshop on Consuming Linked Data, COLD 2013, Sydney, Australia, October 22, (2013).
Rowe, N., The State of Master Data Management, Building the Foundation for a Better Enterprise, (2012).
Ryutov, T., Kichkaylo, T., Neches, R., Access Control Policies for Semantic Networks. In 2009 IEEE International Symposium on Policies for Distributed Systems and Networks, p. 150 -157, (2009).
Schutz, A., Buitelaar, P., RelExt: A tool for relation extraction from text in ontology extension, Proceedings of the 4th International Semantic Web Conference, (2005).
Shadbolt, et al., Linked open government data: lessons from Data.gov.uk. IEEE Intelligent Systems, 27, (3), Spring Issue, 16-24, (2012).
Thomson Reuters Technical Report, ORCID: The importance of proper identification and attribution across the scientific literature ecosystem, (2013).
Tuchinda, R., Szekely, P., and Knoblock, C. A., Building Data Integration Queries by Demonstration, In Proceedings of the International Conference on Intelligent User Interface, (2007).
Tuchinda, R., Knoblock, C. A., Szekely, P., Building Mashups by Demonstration, ACM Transactions on the Web (TWEB), 5(3), (2011).
Ul Hassan, U. U., O'Riain, S., Curry, E., Towards Expertise Modelling for Routing Data Cleaning Tasks within a Community of Knowledge Workers. *Proceedings of the 17th International Conference on Information Quality,* (2012).
Wise, J., The Pistoia Alliance, Collaborate to Innovate: Open Innovation, SMEs, motives for and barriers to cooperation, (2011).

Annex 1. Use Case Analysis

This section provides a classification of the sector case studies according to different dimensions of analysis. Specific case studies are included on the analysis according to the availability of the information.Table 11-1 *classifies the data sources of the case study according to the features of the curated data.*

	quantity	timeframe/ lifespan	key records/ metadata	ownership	access/use
NYT	10^7 / articles	complete/ undetermined	articles, associated categories	partially open	internal annotation platform and public API for entities
PDB	10^3 / protein structures	complete/ undetermined	protein structures, associated protein data	Open	public search and navigation interface, data download
ChemSpider	10^3/ molecules	complete/ undetermined	molecule data, associated links	Open	public search and navigation interface, data download
Wikipedia	10^6 / articles	complete/ undetermined	articles, links	Open	public search and navigation interface, data download
legislation.gov.uk	10^3/ legislation items	complete/ undetermined	laws, categories, historical differences and evolution	Open	public search and navigation interface

Table 11-1 Data features associated with the curated data

Table 11-2 describes the data quality dimensions which are critical for each case study. An analysis of the categories in the data curation case studies shows that projects curating smaller and more structured datasets tend to assess *completeness* as a critical dimension. *Discoverability, accessibility and availability* as well as *interpretability and reusability* are critical dimensions for all of the projects, which are data providers for either third-party data consumers or to large number of data consumers. *Accuracy* tends to be a central concern for projects curating structured data.

Data quality	Discoverability, Accessibility & Availability	Completeness	Interpretability & Reusability	Accuracy	Consistency & Integrity	Trust worthiness	Timeliness
NYT	High	Medium	High	Medium	Medium	High	Low
PDB	High	High	High	High	High	High	High
Chemspider	High	High	High	High	High	High	High
Wikipedia	High	Medium	High	Medium	Medium	High	High
legislation.gov.uk	High	High	High	High	High	High	High

Table 11-2: Critical data quality dimensions for existing data curation projects

Table 11-3 compares the relevance of the data curation roles in existing projects. There is a large variability on the distribution of roles across different data curation infrastructures. The larger the scale (NYT, Wikipedia case studies), the more open is the participation on the curation process. Also the more structured is the data (ChemSpider), the larger is the participation and importance of the *rules manager* and *data validator* roles. Most of the projects have data consumers which are domain experts and also work as curators. However for open Web projects the ratio between data consumers and active curators is very high (for larger Wikipedia versions, around 0.02-0.03%).[1]

	Coordinator	Rules Mng.	Schema Mng.	Data Validator	Domain Expert	Consumers as curators	# of curators
NYT	High	High	High	High	High	High	10^3
PDB	Medium	Medium	High	High	High	Medium	N/A
ChemSpider	High	Medium	Medium	High	High	Medium	10
Wikipedia	High	High	Low	High	High	Medium	10^3
legislation.gov.uk	Low	Low	High	Low	High	Low	N/A

Table 11-3: Existing data curation roles and their coverage on existing projects.

Table 11-4 describes the coverage of each core technological dimension for the exemplar curation platforms. More principled *provenance capture and management* is still underdeveloped in existing projects and it is a core concern among most of them (Buneman et al., 2006). Since most of the evaluated projects target open data, *permissions and access management* are modelled in a coarse grained manner, covering the existing curation roles and not focusing on specific data-item level access. As data curation moves in the direction of private collaboration networks, this dimension should get greater importance. *Data consumption and access infrastructures* usually rely on traditional Web interfaces (keyword search and link navigation). More advanced semantic search, query and interaction (Freitas et al., 2012) capabilities are still missing from most of the existing systems and impact both on the consumption and on the data curation tasks (such as in the schema matching). For projects curating structured and semi-structured data, interoperability on the data model and on the conceptual model levels are recognised as one of the most critical features, pointing in the direction on the use of standardized data representation for data and metadata.

[1] Wikimedia users: http://strategy.wikimedia.org/wiki/Wikimedia_users, last access September 2013.

Technological/ Infrastructure Dimensions	Data Representation	Data Transformation/ Integration Approaches	Data Consumption/ Access Infrastructure	Infrastructures for Human Collaboration and Validation	Provenance and Trust Management	Permission/Access Management
NYT	Domain Specific/ Linked Data	Human	Navigation, Search	Critical	Low	Coarse-grained
PDB	Domain-Specific	N/A	Search Interface	Data deposits	Low	Coarse-grained
ChemSpider	Relational(Moving to RDF)	Human and Algorithmic	Search Interface, Navigation	Critical	Low	Medium
Wikipedia	HTML	Human and Algorithmic	Navigation, Search	Critical	Medium	Medium
legislation.gov.uk	XML/RDF	Human	Navigation, Search	Critical	Low	Coarse-grained

Table 11-4: Technological infrastructure dimensions

Table 11-5 provides some of the key features from the case studies.

Case study	Consumers	Semantic Technologies	Consumers as curators	Data Validation
ChemSpider	chemists	Yes	Yes	Yes
Protein Data Bank	biomedical community	Yes	Yes	Yes
Press Association	public	Yes	No	Yes
Thomson Reuters	public	Yes	No	Yes
The New York Times	public	Yes	Yes	Yes

Table 11-5: Summary of sector case studies