Chapter 6
Big Data Curation

André Freitas and Edward Curry

6.1 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 organizations (Curry et al. 2010).

The emergence of new platforms for decentralized 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 organizations (Brodie and 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 of 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.

A. Freitas (✉) • E. Curry
Insight Centre for Data Analytics, National University of Ireland Galway, Lower Dangan, Galway, Ireland
e-mail: andre.freitas@insight-centre.org; edward.curry@insight-centre.org

© The Author(s) 2016

J.M. Cavanillas et al. (eds.), *New Horizons for a Data-Driven Economy*,
DOI 10.1007/978-3-319-21569-3_6

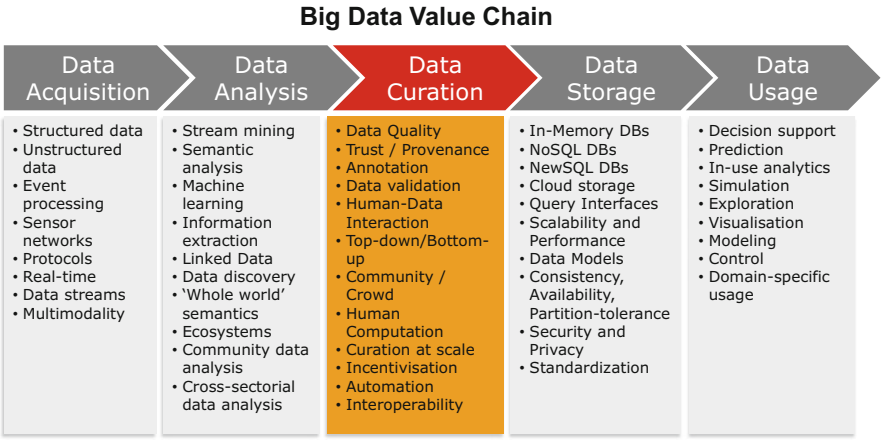

Fig. 6.1 Data curation in the big data value chain

The position of big data curation within the overall big data value chain can be seen in Fig. 6.1. 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 *technological* dimensions. This chapter analyses the data dynamics in which data curation is inserted, investigates future requirements and emerging trends for data curation, and briefly describes exemplar case studies.

6.2 Key Insights for Big Data Curation

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, organizations 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 organize 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 organizations. More complete models support a larger number of answers through data analysis. Data curation practices and technologies will progressively become more present in contemporary data management environments, facilitating organizations 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 organizations*.

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 are 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 reduces 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 and 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.

There is the 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 is required. 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 to industry environments. Many NLP and ML tools have uncertainty levels associated with their results and are dependent on training over large 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.

6.3 Emerging Requirements for Big Data Curation

Many big data scenarios are associated with 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 in statements such as: “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 2014).

Reusing data that 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 mind-set 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" (Kevin Ashley 2014). 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, methodologies are needed to describe data so that it is actually reusable outside its original context (Kevin Ashley 2014). 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" (Kevin Ashley 2014). 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 "*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*". 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*".

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 and Strong 1996; Knight and 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.

Despite the fact that data heterogeneity and data quality were concerns already present before the big data scale era (Wang and Strong 1996; Knight and Burn 2005), 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* (Curry and Freitas 2014). Traditional relational data management

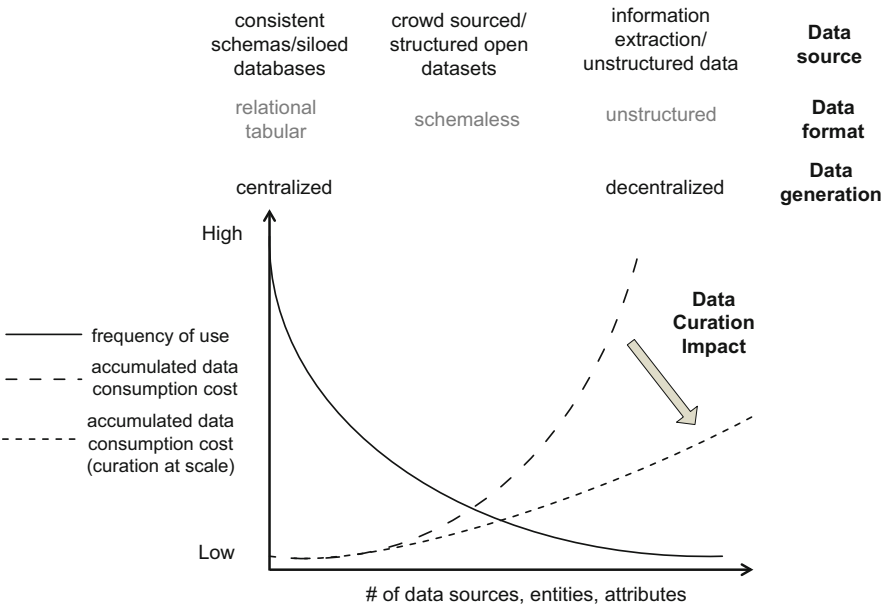

Fig. 6.2 The long tail of data curation and the scalability of data curation activities

environments were focused on data that mapped to frequent business processes and were regular enough to fit into a relational model. The *long tail of data variety* (see Fig. 6.2) expresses the shift towards expanding the data coverage of data management environments towards data that 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* (Fig. 6.2), allowing data curation tasks to be addressed under limited time constraints.

Scaling up data curation is a multidisciplinary problem that requires the development of *economic models*, *social structures, incentive models,* 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.

6.4 Social and Economic Impact of Big Data Curation

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 reality. While some big data use cases are based

on large scale but small schema and 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 a 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 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 that 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 to create 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).

In 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 2014). For pharmaceutical companies data curation is central to organizational 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 improvement of data curation tools and methods directly provides greater efficiency of the knowledge discovery process, maximizes return of investment per data item through reuse, and improves organizational transparency.

6.5 Big Data Curation State of the Art

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

Master Data Management is composed of the processes and tools that support a single point of reference for the data of an organization, an authoritative data source. Master Data Management (MDM) tools can be used to remove duplicates and standardize data syntax, as an authoritative source of master data. 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, and semantic enrichment.

Master data management is highly associated with data quality. According to Morris and 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 systems

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

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 categorization and normalization 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 and Blanke (2012):

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

Crowdsourcing 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 a 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 where the 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 and von Ahn 2011).

General-purpose crowdsourcing service platforms such as CrowdFlower (CrowdFlower Whitepaper 2012) or Amazon Mechanical Turk (Ipeirotis 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).

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 platforms), Wikis have proven to scale to very large user bases. As of 2014, Wikipedia counted more than 4,000,000 articles and has a community with more than 130,000 active registered contributors.

[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/

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 et al. 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.

6.5.1 Data Curation Platforms

- **Data Tamer:** 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:** 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:** This database system answers SQL 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:** 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 recommends flagged articles to a Wikipedia editor based on their profile.

6.6 Future Requirements and Emerging Trends for Big Data Curation

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.

6.6.1 Future Requirements for Big Data Curation

The list of future requirements was compiled by selecting and categorizing the most recurrent demands in a state-of-the-art survey and which emerged in domain expert interviews as a fundamental direction for the future of data curation. 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 and entertainment, telco, manufacturing), and environmental.
- **Priority:** Covers the level of priority that is associated with the requirement. Possible values are: short-term (<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, non-governmental organizations, and user communities.

6.6.2 Emerging Paradigms for Big Data Curation

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. These approaches are summarized in Table 6.2.

Table. 6.1 Future requirements for data curation

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 and 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 6.2 Emerging approaches for addressing the future requirements

Requirement category	Emerging approach	Adoption/status	Exemplar use case
Incentives creation and social engagement mechanisms	Open and interoperable data policies	Early-stage/Limited adoption	Data.gov.uk
	Better recognition of the data curation role	Lacking adoption/ Despite the exemplar use cases, the data curator role is still not recognized	Chemspider, Wikipedia, Protein Data Bank
	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)
Economic models	Pre-competitive partnerships	Seminal use cases	Pistoia Alliance (Barnes et al. 2009)
	Public–private partnerships	Seminal use cases	Geoconnections (Harper 2012)
	Quantification of the economic impact of data	Seminal use cases	Technopolis Group (2011) ("Data centres: their use, value and impact")
Curation at scale	Human computation and 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 and 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 ease of interactivity	Seminal tools available	OpenRefine
	Natural language interfaces, schema-agnostic queries	Research stage	IBM Watson (Ferrucci et al. 2010), Treo (Freitas and Curry 2014)

(continued)

Table 6.2 (continued)

Requirement category	Emerging approach	Adoption/status	Exemplar use case
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 and Atluri (2003), Ryutov et al. (2009), Kirrane et al. (2013), Rodrıguez-Doncel et al. (2013)
Standardization and interoperability	Standardized data model	Standards are available	RDF(S), OWL
	Reuse of vocabularies	Technologies for supporting 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 and Le Novère 2007)
	Nanopublications	Emerging concept	Mons and Velterop (2009), Groth et al. (2010)
	Investigation of theoretical principles and domain-specific models for data curation	Emerging concept	Pearl and 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)

6.6.2.1 Social Incentives and Engagement Mechanisms

Open and Interoperable Data Policies 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 to make 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 of 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 be better recognized across the scientific and publishing pipeline. Some organizations and projects have already a clear definition of different data curator roles. Examples are Wikipedia, New York Times (Curry et al. 2010), and Chemspider (Pence and Williams 2010). 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 that is meaningful for a user community. Projects such as Wikipedia, 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 and social or professional status, social impact, meaningfulness, or fun. The understanding of these principles and the development of the mechanisms behind the engagement of large user bases is an important issue for amplifying data curation efforts.

6.6.2.2 Economic Models

Emerging economic models 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 organizations, which are typically competitors, collaborate in parts of the Research & Development (R&D) process which does not impact on 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 (Barnes et al. 2009), 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. The Pistoia Alliance was founded by pharmaceutical companies such as AstraZeneca, GSK, Pfizer, and Novartis, and examples of shared resources include data and data infrastructure tools.

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. 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. GeoConnections Canada is an example of a 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 2014). GeoConnections has been developed on a collaborative model involving the participation of federal, provincial, and territorial agencies, and the private and academic sectors.

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*" (Technopolis Group 2011), which provides a quantitative account of the value creation process of eight data centres. The creation of quantitative financial measures can provide the required evidence to support data infrastructure investments both public and private, creating sustainable business models grounded on data assets, expanding the existing data economy.

6.6.2.3 Curation at Scale

Human Computation and Crowdsourcing Services Crowdsourcing platforms are rapidly evolving but there is still a major opportunity for *market differentiation* and *growth*. CrowdFlower, for example, is evolving in the direction of *providing better APIs*, supporting *better integration with external systems*.

Within 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 (Hassan et al. 2012). Furthermore combining work by different people on the same task might also help in improving the quality of work (Law and von Ahn 2009). Recruitment of suitable humans for computation is a major challenge of human computation.

Today, these platforms are mostly restricted to tasks that can be delegated to a paid generic audience. Possible future differentiation avenues include: (1) support for highly specialized domain experts, (2) more flexibility in the selection of demographic profiles, (3) creation of longer term (more persistent) relationships with teams of workers, (4) creation of a major general purpose open crowdsourcing service platform for voluntary work, and (5) using historical data to provide more productivity and automation for data curators (Kittur et al. 2007).

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 2014). 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: (1) capture and representation of user actions; (2) annotation mechanisms/vocabulary reuse; (3) ability to handle large-scale data; (4) better search capabilities; and (5) integration with multiple data sources.

Extending applications with large user bases for data curation provides 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 2014).

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 core data curation functionalities into a large-scale data curation platform is a fundamental element for organizations that do large-scale data curation. Platforms such as Open Refine[4] and Karma (Gil

[4] http://openrefine.org/

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. ETL pipelines can be seen as concentrating recurrent activities that become 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 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).

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 of providing an algorithmic approach for the validation or annotation of the data against reference standards (Data Curation Interview: Antony Williams 2014). 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 2014). 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 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. There is significant potential for the application of machine learning in the data curation field. Two research areas that can 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 and Schmid 2008; Lieberman 2001) is a set of end user development approaches in which 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 with research on PbD data integration (Tuchinda et al. 2007, 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 a system that uses at its centre a statistical model to determine the probability of an answer being correct (Ferrucci et al. 2010). Uncertainty models can also be used to route tasks according to the level of expertise, minimizing the cost and maximizing the quality of data curation.

6.6.2.4 Human–Data Interaction

Interactivity and Ease of Curation Actions Data interaction approaches that 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 2014). Therefore, appropriate summarization and visualization of data is important not only from the usage perspective but also from the maintenance perspective (Hey and Trefethen 2004). Specifically, for the collaborative methods of data cleaning, it is fundamental to enable the 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**[5] **(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**[6]**:** 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.

[5] http://d3js.org/

[6] http://www.tableausoftware.com/public/

- **Open Refine**[7]**:** This open source application allows users to clean and transform data from a variety of formats such as CSV, XML, RDF, JSON, etc. Open Refine is particularly useful for finding outliers in data and checking the distribution of values in columns through facets. It allows data reconciliation with external data sources such as Freebase and OpenCorporates.[8]

Structured query languages such as SQL are the default approach for interacting with databases, together with graphical user interfaces that 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 that 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. 2012a, b; Kaufmann and Bernstein 2007). Flexible approaches for database querying depend 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.

6.6.2.5 Trust

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[9] 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 2014). However, there is still a relatively low adoption on provenance

[7] https://github.com/OpenRefine/OpenRefine/wiki

[8] https://www.opencorporates.com

[9] http://www.w3.org/TR/prov-primer/

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 groups of users 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 and 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).

6.6.2.6 Standardization and 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)[10] facilitate 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 New York Times (NYT) Index (Curry et al. 2010) or the Protein Data Bank (PDB) (Bernstein et al. 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 that 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 and Euzenat 2005; Freitas et al. 2012a, b) are central for reducing the burden of manual vocabulary mapping on the end user side, reducing the burden for terminological reuse (Freitas et al. 2012a, b).

[10] http://www.w3.org/TR/rdf11-primer/

Improved Integration and Communication between Curation Tools Data is created and curated in different contexts and using different tools (which are specialized 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 specialized tools to make the final result trustworthy (Data Curation Interview: Paul Groth 2014; Data Curation Interview: James Cheney 2014). The creation of standardized 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.

6.6.2.7 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 still need to be better formalized. The adoption of methods such as minimum information models (La Novere et al. 2005) 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 and 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 and Velterop 2009; Groth et al. 2010). Nanopublications are core scientific statements with associated contexts (Groth et al. 2010), which aim 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 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 and Bareinboim 2011) are examples of theoretical models that can impact data curation, guiding users towards the generation and representation of data that can be reused in broader contexts.

6.6.2.8 Unstructured and 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* (Fig. 6.2).

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. 2012a, b; Schutz and Buitelaar 2005; Han et al. 2011; Data Curation Interview: Helen Lippell 2014). 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 2014).

Use of Open Data to Integrate Structured and 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 organizational 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 that are built from large-scale collections (Freitas et al. 2012a, b), can be applied to structured databases (Freitas and Curry 2014) and are examples of approaches that can be used to enrich the semantics of the data.

Natural Language Processing Pipelines The Natural Language Processing (NLP) community has mature approaches and tools that can be directly applied to projects that deal with unstructured data. Open source projects such as Apache UIMA[11] facilitate the integration of NLP functionalities into other systems. Additionally, strong industry use cases such as IBM Watson (Ferrucci et al. 2010), Thomson Reuters, The New York Times (Curry et al. 2010), and the Press

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

Association (Data Curation Interview: Hellen Lippell) are shifting the perception of NLP techniques from the academic to the industrial field.

6.7 Sectors Case Studies for Big Data Curation

In this section, case studies are discussed 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.

6.7.1 Health and Life Sciences

6.7.1.1 ChemSpider

ChemSpider[12] is a search engine that provides free 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 with 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. Used by chemists for identifier conversion and 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 of the manual annotation and correction of data (Pence and Williams 2010). This may include changes to the chemical structures of a compound, addition or deletion of identifiers, associating links between a chemical compound, 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 highlight the need for appropriate action by a master curator.
- As a registered member with curation rights, directly curate the data or 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

[12] http://www.chemspider.com

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 levels 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 to determine whether or not they are validly represented, or if 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, flag it, and delete it. ChemSpider is in the process of implementing an automated recognition system that will measure the contribution effort of curators through the data validation and engagement process. The contribution metrics can be publicly viewable and accessible through a central profile for the data curator.

6.7.1.2 Protein Data Bank

The Research Collaboratory for Structural Bioinformatics Protein Data Bank[13] (RCSB PDB) is a group dedicated to improve the understanding of the functions of biological systems through the study of 3D structure of biological macromolecules. The PDB has had over 300 million dataset downloads.

A significant amount of the curation process at PDB consists of providing standardized vocabulary for describing the relationships between biological entities, varying from organ tissue to the description of the molecular structure. The use of standardized vocabularies helps with the nomenclature used to describe protein and small molecule names and their descriptors present in the structure entry. The data curation process covers the identification and correction of inconsistencies over the 3D protein structure and experimental data. In order to implement a global hierarchical governance approach to the data curation workflow, PDB 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 datasets are released on a weekly basis to keep all sources consistent with the current standards and to ensure good data curation quality.

[13] http://www.pdb.org

6.7.1.3 FoldIt

Foldit (Good and Su 2011) is a popular example of a 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 that 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 overcome this problem, Foldit uses competitive protein folding to generate the best proteins (Eiben et al. 2012).

6.7.2 Media and Entertainment

6.7.2.1 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 (text, data, photos, and videos) to 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 ways 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 PA 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 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 Fig. 6.3.

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 metadata manager then reviews the classification and the content is published online.

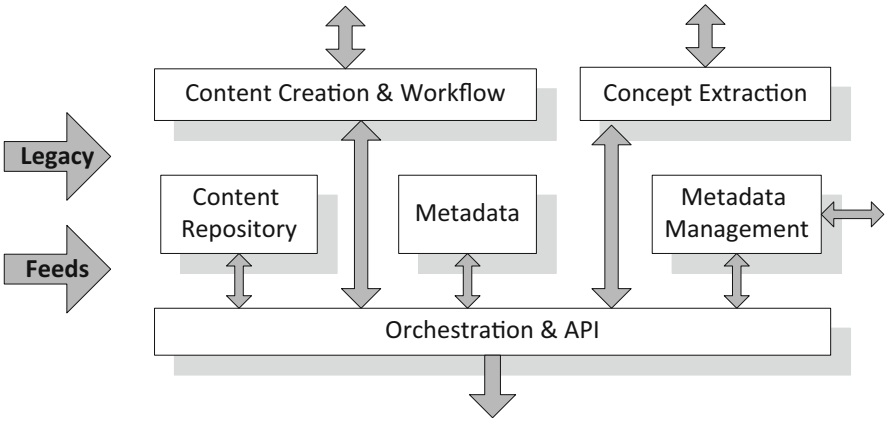

Fig. 6.3 Press Association content and metadata pattern workflow

6.7.2.2 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).

The New York Times' curation pipeline (see Fig. 6.4) starts with an article getting out of the newsroom. The first level curation consists of 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[14]). Teragram uses a set of linguistic extraction rules, which are created by the taxonomy managers based on a subset of the controlled vocabulary used by the Index Department. Teragram suggests tags based on the index vocabulary that can potentially describe the content of the article (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.

6.7.3 Retail

6.7.3.1 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

[14] SAS Teragram http://www.teragram.com

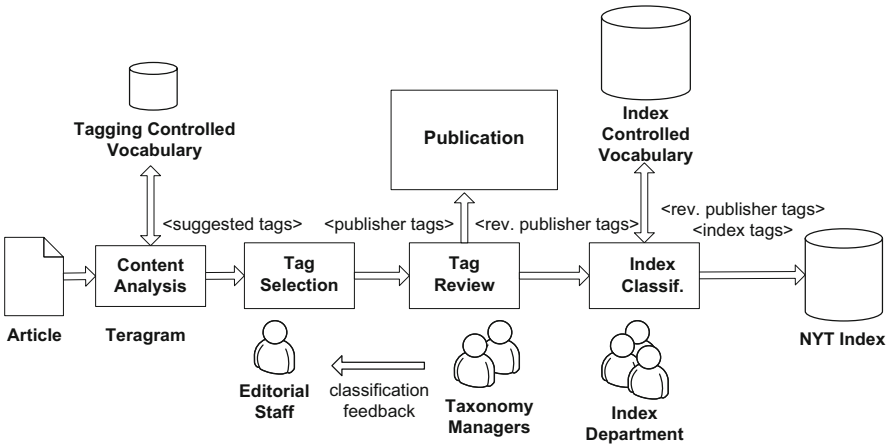

Fig. 6.4 The New York Times article classification curation workflow

important issues of data quality: managing product taxonomies and finding identifiers in product descriptions. Crowdsourced workers help eBay in improving the speed and quality of product classification algorithms at lower costs.

6.7.3.2 Unilever

Unilever is one of the world's largest manufacturers of consumer goods, with global operations. Unilever utilized crowdsourced human computation within their marketing strategy for new products. Human computation was used to gather sufficient data about customer feedback and to analyse public sentiment of social media. Initially Unilever developed a set of machine-learning algorithms to conduct an 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, Unilever 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.

6.8 Conclusions

With the growth in the number of data sources and of decentralized content generation, ensuring data quality becomes a fundamental issue for 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 investigation that included a comprehensive literature analysis, survey, interviews with data curation experts,

questionnaires, and case studies, the future requirements and emerging trends for data curation were identified. The 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.

Open Access This chapter is distributed under the terms of the Creative Commons Attribution-Noncommercial 2.5 License (http://creativecommons.org/licenses/by-nc/2.5/) which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

The images or other third party material in this book are included in the work's Creative Commons license, unless indicated otherwise in the credit line; if such material is not included in the work's Creative Commons license and the respective action is not permitted by statutory regulation, users will need to obtain permission from the license holder to duplicate, adapt, or reproduce the material.

References

Ashley, K. (2014). Data curation interview. Expert interview series for the EU-project BIG (318062; ICT-2011.4.4). http://big-project.eu/text-interviews

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). 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* (pp. 722–735).

Barnes, M. R., Harland, L., Foord, S. M., Hall, M. D., Dix, I., Thomas, S., et al. (2009). Lowering industry firewalls: Pre-competitive informatics initiatives in drug discovery. *Nature Reviews Drug Discovery, 8*(9), 701–708.

Berners-Lee, T. (2009). Linked data design issues. http://www.w3.org/DesignIssues/LinkedData.html

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr., Brice, M. D., & Rodgers, J. R. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology, 112*(3), 535–542.

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

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

Buneman, P., Cheney, J., Tan, W., & Vansummeren, S. (2008). Curated databases. In *Proceedings of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*.

Cheney, J. (2010). Causality and the semantics of provenance. *arXiv preprint arXiv:1004.3241*.

Cheney, J. (2014). Data curation interview. Expert interview series for the EU-project BIG (318062; ICT-2011.4.4). http://big-project.eu/text-interviews

Cragin, M., Heidorn, P., Palmer, C. L., & Smith, L. C. (2007). *An educational program on data curation, ALA science & technology section conference*.

CrowdFlower. (2012). *Crowdsourcing: Utilizing the cloud-based workforce* (Whitepaper).

Curry, E., & Freitas, A. (2014). *Coping with the long tail of data variety*. Athens: European Data Forum.

Curry, E., Freitas, A., & O'Riáin, S. (2010). The role of community-driven data curation for enterprise. In D. Wood (Ed.), *Linking enterprise data* (pp. 25–47). Boston, MA: Springer US.

Cypher, A. (1993). *Watch what i do: Programming by demonstration*. Cambridge, MA: MIT Press.

Doan, A., Ramakrishnan, R., & Halevy, A. (2011). Crowdsourcing systems on the world-wide web. *Communications of the ACM, 54*(4), 86–96.

Eiben, C. B., et al. (2012). Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nature Biotechnology, 30*, 190–192.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., et al. (2010). Building Watson: An overview of the DeepQA project. *AI Magazine, 31*(3), 59–79.

Flener, P., & Schmid, U. (2008). An introduction to inductive programming. *Artificial Intelligence Review, 29*, 45–62.

Fortson, L., Masters, K., Nichol, R., Borne, K., Edmondson, E., & Lintott, C., et al. (2011). *Galaxy Zoo: Morphological classification and citizen science, machine learning and mining for astronomy*. Chapman & Hall.

Franklin, M., Halevy, A., & Maier, D. (2005). From databases to dataspaces: A new abstraction for information management. *ACM SIGMOD Record, 34*(4), 27–33.

Freitas, A., Carvalho, D., Pereira da Silva, J. C., O'Riain, S., & Curry, E. (2012a). 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)*.

Freitas, A., & Curry, E. (2014). 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.

Freitas, A., Curry, E., Oliveira, J. G., & O'Riain, S. (2012b). Querying heterogeneous datasets on the linked data web: Challenges, approaches and trends. *IEEE Internet Computing, 16*(1), 24–33.

Freitas, A., Oliveira, J. G., O'Riain, S., Curry, E., & Pereira da Silva, J. C. (2011). 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)*.

Gartner. (2007). 'Dirty Data' is a Business Problem, Not an IT Problem, says Gartner, Press release.

Gil, Y., Szekely, P., Villamizar, S., Harmon, T. C., Ratnakar, V., Gupta, S., et al. (2011). Mind your metadata: Exploiting semantics for configuration, adaptation, and provenance in scientific workflows. In *Proceedings of the 10th International Semantic Web Conference (ISWC)*.

Goble, C. (2014). Data curation interview. Expert interview series for the EU-project BIG (318062; ICT-2011.4.4). http://big-project.eu/text-interviews

Good, B. M., & Su, A. I. (2011). Games with a scientific purpose. *Genome Biology, 12*(12), 135.

Groth, P., Gibson, A., & Velterop, J. (2010). The anatomy of a nanopublication. *Information Services and Use, 30*, 1–2. 51–56.

Groth, P. (2014). Data curation interview. Expert interview series for the EU-project BIG (318062; ICT-2011.4.4). http://big-project.eu/text-interviews

Han, X., Sun, L., & Zhao, J. (2011). Collective entity linking in web text: A graph-based method. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Harper, D. (2012). *GeoConnections and the Canadian Geospatial Data Infrastructure (CGDI): An SDI Success Story, Global Geospatial Conference*.

Hassan, U. U., O'Riain, S., & Curry, E. (2012). Towards expertise modelling for routing data cleaning tasks within a community of knowledge workers. In *Proceedings of the 17th International Conference on Information Quality*.

Hedges, M., & Blanke, T. (2012). Sheer curation for experimental data and provenance. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries* (405–406).

Hey, T., & Trefethen, A. E. (2004). UK e-science programme: Next generation grid applications. *International Journal of High Performance Computing Applications, 18*(3), 285–291.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., & Yon Rhee, S. (2008). Big data: The future of biocuration. *Nature, 455*(7209), 47–50.

Ipeirotis, P. G. (2010). Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students, 17*(2), 16–21.

Kaggle. (2005). *Go from big data to big analytics*. http://www.kaggle.com/

Kaufmann, E., & Bernstein, A. (2007). How useful are natural language interfaces to the semantic web for casual end-users? In *Proceedings of the 6th International The Semantic Web Conference* (pp. 281–294).

Khatib, F., DiMaio, F., Foldit Contenders Group, Foldit Void Crushers Group, Cooper, S., Kazmierczyk, M. et al. (2011). Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature Structural and Molecular Biology, 18*, 1175–1177.

Kirrane, S., Abdelrahman, A., Mileo, S., & Decker, S. (2013). Secure manipulation of linked data. In *Proceedings of the 12th International Semantic Web Conference*.

Kittur, A., Chi, E., Pendleton, B. A., Suh, B., & Mytkowicz, T. (2007). Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World Wide Web, 1*(2), 19.

Knight, S. A., & Burn, J. (2005). Developing a framework for assessing information quality on the World Wide Web. *Informing Science, 8*, 159–172.

Kong, N., Hanrahan, B., Weksteen, T., Convertino, G., & Chi, E. H. (2011). VisualWikiCurator: Human and machine intelligence for organizing wiki content. In *Proceedings of the 16th International Conference on Intelligent User Interfaces* (pp. 367–370).

La Novere, N., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., et al. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnology, 23*(12), 1509–1515.

Laibe, C., & Le Novère, N. (2007). MIRIAM resources: Tools to generate and resolve robust cross-references in Systems Biology. *BMC Systems Biology, 1*, 58.

Law, E., & von Ahn, L. (2009). Input-agreement: A new mechanism for collecting data using human computation games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (vol. 4, pp. 1197–1206).

Law, E., & von Ahn, L. (2011). Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 5*, 1–121.

Lieberman, H. (2001). *Your wish is my command: Programming By example*. San Francisco, CA: Morgan Kaufmann.

Lippell, H. (2014). Data curation interview. Expert interview series for the EU-project BIG (318062; ICT-2011.4.4). http://big-project.eu/text-interviews

Lord, P., Macdonald, A., Lyon, L., & Giaretta, D. (2004, September). From data deluge to data curation. In *Proceedings of the UK e-science all hands meeting* (pp. 371–357).

Lynch, N. (2014). Data curation interview. Expert interview series for the EU-project BIG (318062; ICT-2011.4.4). http://big-project.eu/text-interviews

Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011, September). DBpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems* (pp. 1–8). New York: ACM.

Mons, B., Ashburner, M., Chichester, C., van Mulligen, E., Weeber, M., den Dunnen, J., et al. (2008). Calling on a million minds for community annotation in WikiProteins. *Genome Biology, 9*(5), R89.

Mons, B., & Velterop, J. (2009). *Nano-Publication in the e-science era, International Semantic Web Conference*.

Morris, H. D., & Vesset, D. (2005). *Managing Master Data for Business Performance Management: The Issues and Hyperion's Solution, Technical Report*.

Norris, R. P. (2007). How to make the dream come true: The astronomers' data manifesto. *Data Science Journal, 6*, S116–S124.

Palmer, C. L., et al. (2013). *Foundations of Data Curation: The Pedagogy and Practice of "Purposeful Work" with Research Data*.

Pearl, J., & Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In *Proceedings of the 25th National Conference on Artificial Intelligence (AAAI)*.

Pence, H. E., & Williams, A. (2010). ChemSpider: An online chemical information resource. *Journal of Chemical Education, 87*(11), 1123–1124.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. http://altmetrics.org/manifesto/

Qin, L., & Atluri, V. (2003). Concept-level access control for the Semantic Web. In *Proceedings of the ACM Workshop on XML Security – XMLSEC '03*. ACM Press.

Rodrıguez-Doncel, V., Gomez-Perez, A., & Mihindukulasooriya, N. (2013). 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. (2012). *The state of master data management, building the foundation for a better enterprise*. Aberdeen Group.

Ryutov, T., Kichkaylo, T., & Neches, R. (2009). Access control policies for semantic networks. In *2009 I.E. International Symposium on Policies for Distributed Systems and Networks* (pp. 150–157).

Schutz, A., & Buitelaar, P. (2005). RelExt: A tool for relation extraction from text in ontology extension. In *Proceedings of the 4th International Semantic Web Conference*.

Sewash, J. (2014). Data curation interview. Expert interview series for the EU-project BIG (318062; ICT-2011.4.4). http://big-project.eu/text-interviews

Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W., et al. (2012). Linked open government data: Lessons from Data.gov.uk. *IEEE Intelligent Systems, 27*(3), Spring Issue, 16–24.

Shvaiko, P., & Euzenat, J. (2005). A survey of schema-based matching approaches. *Journal on Data Semantics, IV*, 146–171.

Sheth, A. (1999). 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* (vol. 495, pp. 5–29).

Surowiecki, J. (2005). *The wisdom of crowds*. New York: Random House LLC.

Technopolis Group. (2011). *Data centres: Their use, value and impact* (JISC Report).

Thomson Reuters Technical Report, ORCID: The importance of proper identification and attribution across the scientific literature ecosystem. (2013).

Tuchinda, R., Knoblock, C. A., & Szekely, P. (2011). Building Mashups by demonstration. *ACM Transactions on the Web (TWEB)*, *5*(3), Art. 16.

Tuchinda, R., Szekely, P., & Knoblock, C. A. (2007). Building data integration queries by demonstration. In *Proceedings of the International Conference on Intelligent User Interface*.

Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33.

Williams, A. (2014). Data curation interview. Expert interview series for the EU-project BIG (318062; ICT-2011.4.4). http://big-project.eu/text-interviews