The Daily Shaping of State Transparency: Standards, Machine-Readability and the Configuration of Open Government Data Policies

Samuel Goëta

Telecom ParisTech, Social Sciences Department, France / samuel.goeta@telecom-paristech.fr

Tim Davies

Berkman Klein Center for Internet and Society, UK

Abstract

While many governments are now committed to release Open Government Data under non-proprietary standardized formats, less attention has been given to the actual consequences of these standards for knowledge workers. Unpacking the history of three open data standards (CSV, GTFS, IATI), this paper shows what is actually happening when these standards are enacted in the work practices of bureaucracies. It is built on participant-observer enquiry and interviews focussed on the back rooms of open data, and looking specifically at the invisible work necessary to construct open datasets. It shows that the adoption of open standards is increasingly becoming an indicator of the advancement of open data programmes. Enacting open standards involves much more than simple technical operations, it operates a quiet and localised transformation of bureaucracies, in which the decisions of data workers have substantive consequences for how the open government data and transparency agendas are performed.

Keywords: Open Government Data; Open Standards; Enactment; Infrastructure Studies; Data Assemblages

Introduction

> "It is time for science studies to investigate how data traverse personal, institutional, and disciplinary divides." (Edwards et al., 2011)

The case for using open standards when diffusing online data has been widely discussed for both scientific and government data (Borgman 2007; Robinson et al., 2009; Lathrop & Ruma, 2010). However, little attention has been given to the consequences of these standards for the workers involved in producing and disseminating open data, and for how standards shape the outcomes of data sharing efforts, particularly in the open government domain. Even when standards are introduced into discussions, data is often treated as though it is already available and ready-to-

use, with the actual work required to construct a standardised dataset remaining almost entirely invisible (Bowker, 2000). As the proactive release of government data is increasingly presented as a "superior" mode of delivering government transparency (Birchall, 2014), it becomes vital to ask how data standards are involved in shaping government transparency? Behind the scenes, in the backrooms of open data (Goëta, 2014), what are the consequences of introducing standards for data workers and the actual organisation of government? What impact do decisions made during standardisation have upon the potential uses of open data? By understanding the challenges facing these invisible workers when working with emerging open data standards (Denis & Pontille, 2012), and the way in which standards construct practices both inside and outside the state, we can gain a deeper understanding of how an emphasis on machine-readable data comes to structure ideas and experiences of open government itself.

A growing subject in Science and Technology Studies (STS), data standards are proliferating in the development of large information infrastructures while still remaining largely invisible and taken-for-granted (Star & Ruhleder, 1996; Lampland & Star, 2009; Busch 2011). The numerous studies on open government data that have been conducted to date have largely overlooked how standards shape datasets, what they exclude, and the supplementary burden they require to be implemented. Such an approach is crucial at this particular moment, as many of the standards for an emerging open data infrastructures, embodied in data portals, policy pronouncements and common analysis and visualisation tools are currently being laid down. Rare studies have followed the information infrastructure studies program (Bowker et al., 2010) to understand open government data (Davies, 2012, 2013, 2014) but none has conducted an ethnography of infrastructure (Star, 1999) to understand the implications of these standards in the daily practices of data workers, and the consequences of these standards for the goals of open government. Situated in bureaucracies, our study aims at surfacing the invisible practical work (Suchman, 1995) that supports the implementation of open standards for government data.

In exploring emerging practices of open government data sharing, it is useful to step back to the experience of particular scientific communities over recent decades, where exchanging data has become a crucial matter and datasets are becoming an object of scientific production in their own right (Bowker, 2000; Edwards et al., 2011; Strasser, 2012). As the data required to explore phenomena of interest grows beyond that which any individual researcher or group could collect, distributed scientific collaborations have needed to develop approaches to pool and share data, leading to the creation of vocabularies, schema and markup languages for representing and exchanging data (Zimmerman, 2007, 2008). However, these processes of standardisation are not straightforward or unproblematic. Information infrastructures studies offer a rich framework within which to understand the hidden work going on in order to enable scientists to share data. Edwards (2010) uses the metaphor of "data friction" to describe the efforts required to share data between people and organizations, and it is in response to this friction that many scientific data sharing infrastructures have been developed. Yet this does not necessarily imply that the goal sought should be "frictionless data" (Pollock, 2013). Almklov (2008) finds that standardised data can be experienced by re-users as decontextualised, and difficult to extract meaning from. And several works have shown that metadata, even defined with shared and precise standards, do not lead scientists to reuse data seamlessly, as standards projects have often promised (Edwards et al., 2011; Millerand & Bowker, 2009; Zimmerman, 2008). Recognising science as essentially an open-ended and always unfinished enterprise, Edwards et al. (2011) highlight the importance of considering "metadata-as-process", and paying attention to the social negotiations that go on around data sharing in science, alongside the technical standardisation.

Open Government Data (OGD) is in many ways a younger enterprise than that of open science (Fecher & Friesike, 2014). Since the late 2000s, government across the world have been adopting policies that call for the publication of government held datasets online, in machine-readable forms, and for anyone to re-use without restric-

tion (Yu & Robinson, 2012; Chignard, 2013; Kitchin, 2014). Multiple drivers for this have been cited, from "unlocking" the re-use value of data the state has already paid for to increasing government efficiency, and delivering greater state transparency (Zuiderwijk et al., 2012; Zuiderwijk & Janssen, 2014). As part of a transparency agenda, OGD has been discussed in relation to past regimes of reactive transparency, delivered through Right to Information (RTI) laws, which gave citizens a right to request documents from government (Fumega & Scrollini, 2011; Open Knowledge Foundation, 2011). In RTI, transparency is associated with a clear transaction between a requestor and government, but in OGD, as Peixoto (2013: 203) puts it, public actors can *"characterize transparency as a unilateral act of disclosure"*. For Peixoto (2013: 203), *"transparency may be realized without third parties scrutinizing or engaging with the disclosed information"*, although transparency theorist David Heald quotes Larsson (1998: 40–2) to argue that *"transparency extends beyond openness to embrace simplicity and comprehensibility. For example, it is possible for an organization to be open about its documents and procedures yet not be transparent to relevant audiences if the information is perceived as incoherent"* (Heald, 2006). Within the discourse of OGD, that coherence has come to be defined in terms of machine-readability, and increasingly the adoption of common open standards. OGD advocates have moved from early calls for 'raw data now' (Pollock, 2007; Berners-Lee, 2009), to argue for the adoption of open standards for data publication. Increasingly, efforts have looked to assess the success of open data initiatives with reference to these standards (Cabinet Office, 2013; Atz et al., 2015). Thus, as in scientific collaborations, OGD initiatives are turning towards the construction of new data infrastructures, shaped by the development and deployment of data standards.

Our aim here is thus to understand what is happening when these data standards are actually enacted (Law & Mol, 2008; Millerand & Bowker, 2009) in the work practices of government bureaucracies, and how this impacts upon the construction of state transparency as a component of open government. This paper is built on ethnographically informed participant-observer enquiry in the back rooms of open data: developed iteratively to look at three cases of open data standardisation: from the structuring of diverse data elements to fit with the requirements of a file format specification, through to the mapping of data from internal systems to a rich semantic standard. For each case, we attempt to operate an infrastructural inversion (Bowker, 1994; Bowker & Star, 2000) by looking first at the historical development of particular standards, the work practices that go along with aligning them, the organizational arrangements they create and the way they shape the data the public have access to, and how it can be used. Prior to introducing these standards, we first take a broader look at the role that discourses of standardisation have played in the OGD movement.

Policy and Principles of Open Government Data: Machine-Readability and Open Standards

The Open Government Data movement claims that the proactive publication of the datasets owned by public administration can lead to a new wave of innovation in the use of government data, bringing about a renewal of transparency and a transformation of administrative practices (Janssen et al., 2012). Following the launch in 2009 of the US Data.gov portal, many countries have established policy requirements and legal frameworks for open data, leading to the creation of hundreds of data portals hosting and providing meta-data on a vast spectrum of datasets, provided by national governments, municipalities, international institutions and even some corporations (Web Foundation, 2014). In 2012, G8 member countries signed up to the G8 Open Data Charter, committing to the idea that government data should be 'open by default', and including in an Annexe a list of the kinds of data, from cadastral registers to national budgets, that governments should share (G8, 2013). The G8 Charter has been followed by an International Open Data Charter (2015), which introduces a principle of data 'interoperability', and which, through its technical working group, has been exploring how to recommend data standards for governments to adopt. Within the Open Government Partnership, a voluntary association of over 60 countries committing to increase the availability of information

about government activities, supporting civic participation and improve accountability, action plan commitments to open data have been amongst the most common (Khan & Foti, 2015).

Since the first articulation of common principles for OGD in Sebastopol in 2007 when well-know digital activists such as Lawrence Lessig, Tim O'Reilly, and Aaron Swartz gathered and set out eight key criteria for government data openness, machine readability and open standards have become core claims of the OGD movements. According to these principles, datasets should be provided in "machine-processable" and "non-proprietary" formats (5th and 7th principles). The Sunlight Foundation's (2010) extended "Ten Principles for Open Government Data" place a particular emphasis on the use of "commonly owned" standards, highlighting the importance of standards being freely accessible and fully documented to facilitate their use (Levien, 1998; Russell, 2014), and pointing as well to the process of control over the revision of standards, which, open standards advocates argue, should take place through a predictable, participatory, and meritocratic system (Open Stand, 2012).

This emphasis on machine-readability and open standards can be understood as a reaction against the common publication of government data either in formats such as PDF which present the layout of data, but which frustrate easy digital access to the underlying fields and figures, and the use of file formats that are protected by patents and intellectual property rights, meaning that to read the files requires either proprietary software, or paying license fees for the right and resources to decode and manipulate the data. It is also motivated by a desire to have data files which can be accessed and manipulated in as wide a range of tools as possible, such that even de-jure non-proprietary formats tend to be considered as de-facto closed by developers if established tooling for working with these formats cannot be easily found across a wide range of programming languages and software packages. However, many of the OGD portals in operation around the world still predominantly provide access to files which fail to meet key definitions of machine-readability, and, even if they do, which fail to make use of common standards (Murillo, 2014; Web Foundation, 2015), leading to redoubled efforts to promote 'best practices' for data publication (W3C, 2015). Furthermore, advocates have also been concerned with how data is represented when it is published using machine-readable open formats, looking to also see use of common schemas that define the kinds of fields and values that would be considered valid in a particular kind of data, and which tools reading that data should be able to understand.

Using open standards in releasing government data is now more than a mere principle: it is progressively being required by regulations brought in to implement OGD policies. In 2013, President Obama released a memo, which states that government information must be released under open and machine-readable standards (Obama, 2013). Agencies are required reporting progress on the implementation of open standards 180 days after the memo. The US DATA Act (2014) requires the creation of a common data schema for the exchange of budget information, and the UK Local Government Transparency Code (DCLG, 2014) is accompanied by strong guidance about the fields that should be used for the disclosure of 14 priority datasets (LGA, 2015). Efforts like the International Aid Transparency Initiative, Open Contracting Data Standard and Budget Data Standard are all working to articulate specific standards for open data publication as part of wider political processes seeking to secure sustained information and data disclosure.

However, whilst advocacy for OGD has focussed on 'big tent' arguments suggesting that the provision of open data brings multiple benefits to a diverse range of stakeholders (Weinstein & Goldstein, 2012), critics have presented the open data movement as a tool for marketisation of public services (Bates, 2012) and as the co-option of otherwise radical transparency and civic-technology activism (Bates, 2013). Practitioners in developing country have questioned the assumptions built into standards promoted as global norms. And current practices around open data have also led to concerns that it will "empower the empowered" (Gurstein, 2011) and thus engender regimes of information injustice (Johnson, 2013). Central to this literature is the argument that the open data movement has been defined mostly by

technical considerations, overlooking the political dimensions of the process (Yu & Robinson, 2012; Morozov, 2013) and presuming that the mere provision of data would automatically empower citizens (Gurstein, 2011; McLean, 2011; Donovan, 2012). In particular, Yu and Robinson (2012: 196) denounce the idea that technical criteria, such as the use of open standards in the release of datasets, should be enough to satisfy calls for transparency, writing that: *"An electronic release of the propaganda statements made by North Korea's political leadership, for example, might satisfy all eight of these requirements [Sebastopol principles on Open Government Data], and might not tend to promote any additional transparency or accountability on the part of the notoriously closed and unaccountable regime"*. To these critiques we might also add lessons from science data sharing, to the effect that data standards rarely produce interoperability or interpretability of datasets. Thus any emphasis on machine-readability opens up important conversations about the decisions that are made in constructing data concerning which stakeholders will have their needs prioritised, and how the costs and benefits of adopting standards end up being distributed.

Yet, these critiques noted, the provision of government data under open standards has become a major demand of open data activists. This demand follows a larger history: the Internet protocols were shaped by a discourse on 'openness' of standards. This rhetoric has found a place in a wide variety of movements, asking for software code, hardware, academic publications or governments to be 'opened' to the public by sharing their foundational components (Russell, 2014). However, the demand for 'openness' in standards was not driven only by rhetoric. Open data activists consider that the use of standards facilitates the reuse of data, and gives more specific meaning to demands for machine-readability. But what do these standards and specifications contain? How do they, in practice, ensure or enhance the machine-readability of data? And how does standardised machine-readable data differ from alternative ways data might be shared, shaping in the process who is engaged in open data re-use activity? To address these questions we look in detail at the histories and contemporary implementation of three major standards, used at different levels for opening government data, to understand how they shape both the machine-readability of data, and how they affect wider practices of governmental transparency.

Framework and Methods

Mirroring a common trend in STS research of scholars *"'intervening' while studying science and technology phenomena"* (Karasti et. al., 2016: 4) we enter this field as both practitioners and researchers: involved in initiatives to support open data publication and use practices, whilst also engaged in the scholarly critical study of open data and open government phenomena. Responding to growing discourse on machine-readability and standardisation in the open data field, we sought to identify a series of applications of open data standards in practice, and to apply methods of "infrastructural inversion" to look beyond the surface narratives, and to explore otherwise invisible and ignored work involved in making datasets available as open data.

Three open government data standards are covered by this paper. The first is the CSV (Comma Separated Value) format, which is a general format, used often for tabular or spreadsheet data. The second is specific to the transit field: the GTFS (General Transit Format Specification), offering a schema for transport timetables. The third is the IATI Standard, generated as part of the International Aid Transparency Initiative (IATI), and presenting a schema for detailed disclosure of aid flows. The development of these cases was an iterative process, combining initially independent work from the two authors into a cross-case analysis to draw out key themes and a deeper understanding of the common and divergent labour and impacts implicated in the production of open data according to different standards.

The cases each contribute to understanding different aspects of standardisation. Whilst the broad label 'open data standards' is commonly used to refer to a wide range of different technical artefacts, we note a distinction between standards as *file formats* that enable the exchange of data between systems, without being directly concerned with the semantic contents of the file

and standards as *schema*, which are concerned with describing the fields and data structures a file should contain, seeking to enable the exchange of the meaning of the data as well as the data itself. Both formats and schema, at their respective levels, can be used to perform the technical validation of a data file: determining whether it is structured and encoded according to the file formats specification, and whether it meets validation rules set out within the schema. Although specifying the fields and entities a particular kind of dataset should contain can be done in the abstract, in practice, many schemas are directly related to particular file formats. For example, the GTFS schema assumes a CSV file format, and IATI is based upon XML. From an infrastructural perspective, schema then builds upon the "inertia of the installed base" (Star & Ruhleder, 1996: 113) provided by their chosen file formats, incorporating many of the affordances and constraints that those formats provide.

Data collection itself took place between 2013 and 2015, through a series of interviews and participant-observation activities with 'data workers'. We use the term data worker to capture a wide range of roles within government institutions and their associated agencies. For many of our interviewees, their formal job title was not data related, yet their role has come to involve work in managing or directly producing open datasets. For the CSV and GTFS cases, an initial series of interviews were conducted with project managers in charge of executing an open data policies. They were asked with whom they collaborated for the project to identify the second series of interviews, data producers who have released files in an open data portal. These in-depth interviews were conducted in four French local administrations and in an international institution, each of which had launched some form of open data portal. Following an initial round of analysis drawing out the relationship between file formats and data schema, we introduced a further case drawing on participant-observation and interviews with participants involved in the development and implementation of the International Aid Transparency Initiative (IATI), seeking to explore how far findings from the earlier cases applied outside the French context, and with a different base file format from CSV. Throughout our enquiry we have complemented interview data with examination of data artefacts created in the cases, direct observations of project meetings, document analysis, and an examination of the wider literature related to each of the standards we study.

In the analysis that follows, we start our infrastructural inversion by critically examining the history and institutional context of each standard, and how they have been adopted or promoted within the open data and open government field. We then turn to a synthesis of our empirical data to look at how a number of themes emerging from the research play out across each standard.

Three Standards and Their Stories

Comma-Separated Value

In a nutshell, CSV stands for Comma Separated Values and designates a file format for storing numbers and text in plain-text forms. The format itself is agnostic as to what content the files should contain. It consists of plain text with any number of records, separated by line breaks. In each record, there are fields, which are separated by a character, usually a comma or a tab. All CSV files can be opened in a text editor or a browser, but the data will not be represented as a spreadsheet but rather as simple text. As both humans and machines can read these files as easily as text, they are possible to deal in absence of complete documentation. The CSV format predates personal computers: it has been used since 1967 at least by the IBM programming language Fortran, and has been implemented in virtually all spreadsheet software, and in many data management systems. CSV, easy to work with in most programming languages, makes possible to process data through a simple two-dimensional array of values. In particular, CSV is used for exchanging tabular data between programs and systems.

Although open data activists praise it as a robust standard (Pollock, 2013), only recently have efforts been made to formally standardize CSV. In 2005, Yakov Shafranovich, a software engineer, proposed a Request for Comments (RFC) to the Internet Engineering Task Force (IETF), an organization that develops and promotes the use of open standards on the Internet. Although it is

now categorized as "Informational" by the IETF, RFC 4180 is generally referenced as the de facto standard for the format of a CSV file. In particular, it specifies that the first line should include a header defining each fields, that any field should be quoted with double quotes and that all rows should contain the same numbers of fields. However, the RFC leaves a number of important issues unspecified, which limits the use of CSV for certain users on two particular aspects. First, valid character sets are not defined, but the RFC suggest using the ASCII characters set, a standard known for favouring English-speaking users, rather than the more comprehensive Unicode (Palme & Pargman, 2009). Second, CSV does not specify how to represent particular kinds of values, such as decimal numbers and dates, even though some countries like France use a comma as decimal separator, and countries vary in the date format they use, risking substantial ambiguity in how data entries such as '11/02/2015', for example, should be interpreted.

Further efforts to standardize CSV are ongoing. In particular the W3C (World Wide Web Consortium) has initiated a working group on CSV based on the observation that " a large percentage of the data published on the Web is tabular data, commonly published as comma separated values (CSV) files" (W3C, 2013). The working group was constituted as part of the W3C advocacy for OGD, promoted in particular by its founder Tim Berners-Lee. It is built out of the fact that the format *"is resisted by some publishers because CSV is a much less rich format that can't express important detail that the publishers want to express, such as annotations, the meaning of identifier codes etc."* (W3C, 2013). The ongoing research of the working group will lead to standard metadata that aims to support the automatic interpretation of CSV files on the web, supporting tools to work around the ambiguities of the format: even if CSV files themselves do not become completely standardized.

Many Open Government Data activists praise CSV for its simplicity and its machine-readability, but they also indicate its limits. Tim Berners-Lee (2010) defined a 5-star grading system in which publishing data in CSV with an open license warrants a 3-star grade. The website 5stardata.info[1] indicates that to publish to CSV format *"you might need converters or plug-ins to export the data from the proprietary format"*. The Open Knowledge Foundation (2013) considers it as the *"most simple possible structured format for data […] remaining readable by both machines and humans"* but highlights it is *"not good for data where structure is not especially tabular"*. More recently, the Open Data Institute (2014), also co-founded by Tim Berners-Lee, has declared that 2014 was the *"year of the CSV"*. It declared that it is *"a basic data format that's widely used and deployed […] but it is also the cause of a lot of pain because of inconsistencies in how it is created: CSVs generated from standard spreadsheets and databases as a matter of course use variable encodings, variable quoting of special characters, and variable line endings."* The organization has published a tool called "CSVLint"[2] which tests if a CSV file is "readable" according to a series of rules, enforcing a set of rules for what a CSV file actually should be, drawing on, but going beyond, the basic RFC specification. The tool is based on the observation that *"CSV looks easy, but it can be hard to make a CSV file that other people can read easily"*.

On a practical basis, the limited standardization of CSV means that opening a file in this format can require the user to understand the complexities of encoding data. When opening a CSV file in most spreadsheet software, a box will often open, asking the user to specify which encoding character set is used in the file, as well as the separator character which delimits fields, and the decimal separator. By default, most spreadsheet software will follow the RFC guidelines but in many situations, users will have to manually change the parameters so that the data is displayed as a regular spreadsheet with properly delimited fields. Users commonly accessing data produced on systems with other localisation settings from their own (e.g. in other countries/language communities) are more likely to encounter such prompts. This box adds frictions for the general public in order to use CSV files. While it allows a level of widespread compatibility across the software tools used by developers, it increases practically the complexity of using this format for the everyday task of viewing data in a spreadsheet, and leads to different experiences depending on the user's locality and language.

General Transit Feed Specification

GTFS (General Transit Feed Specification) provides a schema for public transportation schedules oriented towards facilitating the reuse of transit information by software developers. The need for a common standard was driven by the increasing use by commuters of their phones to plan their trips, as well as the success of online digital maps such as Google Maps and OpenStreetMap. Each GTFS "feed" is composed of a series of CSV files compressed in a single ZIP archive. Each file details one aspect of transit information: transit agency, stops, routes, trips, stop times, calendar, special dates, and information on fares or possible transfers. Not all the files are mandatory but the specification requires specific and detailed fields, which should not vary between published files. In contrary to CSV as a standard format, as a standard schema GTFS specifies much more than just the encoding or the layout of the data: it requires transit agencies to transform their data to common structures and to adopt common terms and categories. While both standards tend to ease interoperability of datasets, GTFS requires transit agencies engage in a process of commensuration, adapting their data against shared metrics (Espeland & Stevens, 1998). This process demands considerable resources, and excludes many aspects of reality rendered by the standard as "incommensurable". For example, whilst it may be possible to describe the type of bus running a route within an arbitrary CSV file, within the GTFS schema such additional non-standard columns would be ruled invalid, and effectively meaningless.

The GTFS standard itself was initially developed by a software engineer from Google, Chris Harrelson, in reply to a request from an IT manager of Trimet, the transit agency for the US city of Portland. Harrelson was working on the current Google Transit project, which included public transit timetables in Google Maps. It appears that, through this collaboration with Trimet, the standard closely resembled the data feeds they already had in use. Had the initial collaboration taken place with another locality, it is possible to imagine that GTFS would have looked quite different. After Portland, more than 400 transit operators have now implemented GTFS and publish their data feed with this standard, making GTFS the most widely used open data standard for exchanging transit data. It is published freely with an open source license, and along with the tools necessary to validate a GTFS feed. Google has dropped its brand from the name of the standard but remains active in its development and continues to extend the number of transit feeds usable in Google Maps.

International Aid Transparency Initiative

The International Aid Transparency Initiative (IATI) was launched in 2008 to develop a common approach for aid donors to share information on their projects, budgets and spending. Following wide ranging consultations with aid donor and recipient countries, the project adopted an open data approach, based on the eXtensible Markup Language (XML) data format in 2011, publishing detailed schemas to set out what information should be shared about aid projects and how that information should be represented. Whilst it was initially developed to meet the needs of government aid donors and recipients, the standard is now used by over 400 organisations, including an increasing number of Non-Governmental Organisations.

Unlike CSV (and GTFS), which use a tabular (two-dimensional) data model, the XML format represents data using a tree structure, where data elements can be nested inside other data elements. It also has a range of in-built mechanisms for validating data, defining value types (e.g. date, number etc.), and standardising how multilingual data should be represented. The XML format was developed by a working group at the W3C between 1996 and 1998, and has since gone through a number of iterations. It is derived from Standardised General Markup Language, which has its roots in the mid-1980s, and itself descends from IBM's Generalised Markup Language (GML), which goes back to the 1960s. The particular innovations of XML include better handling of different character encodings (important for exchange of data containing multiple languages), and new approaches to checking the 'well-formedness' of documents as well as their validity against some defined meta-level schemas (Flynn, 2014).

At the core of IATI is a standard for representing records on individual aid activities. These 'iati-

activity' elements can contain project descriptions and classifications, data on project location, budget information, and detailed transaction level reporting of commitments and spending. The standard also allows each activity element to include details of project results, and associated documents. Few elements are made mandatory by the XML schema of the standard, although many are important to have for detailed and forward-looking information on aid. The standard also provides an extensive range of code lists for the classification of activities, some drawn from existing recognised code lists, and others created specifically for, and maintained by, IATI.

In common with many data standards, few aspects of the IATI are completely new. Rather, it was assembled from past precedent, seeking to find a common ground between the existing systems of major aid donors such that it could be at least minimally populated by data already held. The idea of standardised aid information exchange has a long history. Whilst the OECD's Development Assistance Committee Creditor Reporting System (DAC CRS), based on survey data collection of headline statistics from member governments, has been in place since the 1960s, it was in the late 1980s and early 90s that efforts for standardised digital exchange of detailed ongoing project information emerged. The Common Exchange Format for Development Activity Information (CEFDA), a disk-based exchange system, coming before widespread Internet adoption, was the first effort in this direction, although it ultimate saw limited uptake. However, it's field definitions influenced the creation of International Development Markup Language (IDML) in 2001 (Hüsemann, 2001), a format primarily developed to feed data into the Accessible Information on Aid Activities (AiDA) database developed by Development Gateway (initially a World Bank project). IDML and AiDA in turn influenced the development of IATI, both as donors rejected the idea of 'yet-another-database', opting instead for an approach premised on the distributed publication of interoperable data, and as the XML experience of IDML was available to draw upon in building up an IATI standard.

The 'extensible' aspect of XML can also be put to use in IATI, as it allows valid data to embed new fields within the existing structure, declaring alternative 'namespaces' for this data outside of the formal standard. The intent in the IATI case is that this could support de-facto standardisation between small groups of data publishers, without requiring the full process of changing the standard to accommodate use-cases only of concern to a small community of users. However, in practice most extensions to the standard have taken place through the regular revision process, with, for example, more detailed fields for geocoding the location of aid projects recently introduced.

Whilst XML is well suited for exchange of structured data between machines, it can be complex to work with in web applications, and tools exist to help users who are more familiar with tabular data to open and manipulate XML. As a result, IATI has also seen a degree of tool building and secondary standardisation take place, designed to convert the IATI XML data into other formats optimised for different users. A 'data store' has been created which aggregates together known IATI XML files, and then provides various possible CSV rendering of these (each having to choose which elements from the tree-structure of the data to treat as the rows in the file, choosing, for example, between one 'activity' or one 'transaction' per row), and which also offers a JSON (JavaScript Object Notation) format, targeted at web application developers. Each of these alternative formats is in some way 'lossy', containing less information than the XML. Yet, in practice these alternative mediated presentations of the data become the forms that most users are likely to encounter and work with.

Whilst open data standards may often be presented as simple technical artefacts that can be transparency applied to existing datasets, and as a relatively new feature of the open data landscape, these sketches illustrate the long history of even the 'simplest' of standards, and point towards the embedded politics, affordances and limitations of each. We turn then to look at how these standards collide with the work practices of those responsible for making open data available.

The Transformation of Practice: Standards and Data Workers

The use of open standards requires data workers to transform their datasets and to adjust to the standards. This intensive work, led by data producers and open data project managers before opening the data, is rarely measured in advance and is often hidden in the back rooms of open data (Goëta, 2014). As their adoption can require major transformations of pre-existing datasets, standards may increase the complexity of releasing machine-readable and re-usable data. Yet, they can also explicitly or implicitly encode knowledge about how to increase the accessibility of data to a particular community of users.

In order to make a usable CSV dataset, data workers frequently make deep modifications to the original files held by government. The complexity of this transformation is well illustrated by an internal document made by the region Ile-de-France, latter published online as the general guideline of relevance to other organisations releasing open data in CSV format. Entitled "Open data: good practices using Excel"[3], it aims to help data workers publish data in the region's open data portal. The portal policies require data to be published in CSV format, and encourages the geocoding of data entries. Among the recommendations it provides, more than half are directly driven by the specifications of the CSV format. The document asks data workers to fit to the standard, as many aspects of their datasets will simply disappear when changing the format:

> *"One sheet=one dataset": CSV does not support multiple sheets;*
> *"No information should be transmitted by using color—> in CSV format, these data will be suppressed!"*
> *"No merged cells"*
> *"Beware with hidden lines!—> they will display in CSV."*

Besides, the document asks data producers to reorganize the structure of the datasets to fit the RFC specifications of a CSV file which is in use in the region's open data portal:

> *"Column headers on the first lines (=columns titles)";*
> *"No empty cells on columns titles";*
> *"Avoid empty lines or columns";*
> *"Warning with 'orphan' data" designating fields, which are outside of a table and will not display properly in the portal.*

These requirements imply a major transformation of datasets in order to fit it to the CSV standard. The files that officials are being asked to make available under OGD policies were generally not originally produced to be released outside the organization in another format. In the organisations we have studied, it is not the data producers (the subject matter specialists working in the policy areas the data describes) who carry the work of transforming the data to adapt to the CSV format. Instead, open data project managers, whose mission is to actually open the data, take charge of modifying the datasets to fit them into the required formats and standards. In our CSV case, these project managers, originally hired to develop a data portal and foster reuse of the data, have become data managers, directly involved with ensuring the compatibility of specific government datasets with open standards, and interposed between the domain-expert data producers, and the public who access the open data produced. As one explained:

> *"Project manager: When we receive an Excel file, we open it and there are basic stuff such as merged cells, [information in] bold, color…*
> *Interviewer: Do you remove it?*
> *Project manager: Yes, anyway if you want to pass it in CSV, all of a sudden everything disappear and the thing is that for certain files the guys they put color on it although in CSV there is no color. So you have to create other columns.*
> *Interviewer: And how do you do in these cases?*
> *Project manager: Well, you do it manually."*
>
> *(Open data project manager, local authority, France)*

Information erased by the standard has to be rebuilt by the open data project manager who translates this information into a structure that passes the filters of the format or schema.

Creating a GTFS feed also requires intensive work, and a worker to undertake it. Within the organisations we surveyed, transit timetables

are contained in numerous information systems and knowledge about their inner working is spread throughout the division of the organisation. Database managers, existing professionals responsible for various data systems in the organisations, needed to undertake complex work explorating the databases to work out how to actually open their transit timetable data. The exploration is made even more challenging when the data has been released following an externally imposed standard. For GTFS, database managers have to dig throughout the organisation to create a proper feed. One interviewee reported how:

> *"For building our first GTFS feed, we released it with the means available because everyone was not ready in the organisation to publish bus data. So it missed, by the time, around 10% of the stops in the feed. It did not have all the schedules and so, with the feedback from developers, it helped us enhance internally the chain to generate a dataset, to enhance the methods upstream, which create the stops in the different software. After around four or five months of work, we actually succeeded in releasing a dataset, which is actually clean and could be used as such by the developers. […]"*
>
> *(Database manager, transit organisation)*

In this case, it took over four months in this case to create a proper GTFS feed, which contained all the bus stops. A long cry from 'raw data now'. Not only did the GTFS standard require a combination of different databases, but the making of the GTFS feed also required organizational work to align data production between the different data producers. Rather than making a pre-existing dataset transparent, the standard demanded the creation of new infrastructural configurations, and resulted in a new dataset, and a new view on government transit activities that had been unavailable before.

The introduction of the standard and production of externally facing data also surfaced weaknesses in existing internal processes. The same database manager reflecting on data errors describes how:

> *"Sometimes it was a mistake by the system so we fixed it. And sometimes it was just a lack of communication between departments. For instance, we figured out that when a stop changed name, there was not always communication from the person who changed the name. […]"*
>
> *(Database manager, transit organisation)*

As a result, short-term process fixes have been documented, and new practices introduced such as asking everyone changing a bus stop name to e-mails across the organisation to get other systems updated. Longer term, however, the external data standard holds a mirror up to the internal infrastructure, and invites consideration of wider changes. As the database manager reported:

> *"Nowadays, to create a GTFS feed, we mix around 6 or 7 databases. Now it works but we still depend that all the databases are up to date. That's why we are thinking on how to build a new information system for buses in which everything is in one database because now this is really complicated."*
>
> *(Database manager, transit organisation)*

This same dynamic of changing organisational activity was also present in the previously discussed local authority CSV cases. One project manager described his ongoing efforts to ask data producers to structure their datasets according to the standard instead of himself actually transforming the files:

> *"Interviewer: Are you going to transform the data every time they are updated?*
> *Project manager: The thing is we try to educate data producers to well structure their files at the beginning. That will avoid us to effectively remake their files every time. […] We are thinking about it because we figured out that we manage the data at the end of the tunnel."*
>
> *(Open data project manager, local authority, France)*

However, both source data and data standards, are often moving targets. Small changes in source data can disrupt well-established processes for data conversation, such as the first time a non-latin alphabet character occurs in a source dataset,

or when data in an unexpected character encoding is passed into a database system. Schema such as GTFS and IATI are also updated over time, and conventions around CSV on the web continue to evolve. In the case of one large IATI publisher, technical and policy staffs look monthly at ways to follow the evolutions of the standard and maintain interoperability of the datasets with other data providers:

> *"Our journey has been one of continual improvement, to make sure we keep up with the technical standard as it evolves through 1.01, 1.02, 1.03 and so-on. And also continual improvement by recognising where there are problems with the data and fixing them, but also doing that on a very incremental and month-by-month basis. The fact that we publish monthly means you can do that - you can make a minor improvement that makes a big difference and over time they really accumulate. The other aspect of that would be increasing automation. We're fortunately to start off with a fairly automated process in that our entire IATI dataset gets generated nightly, and we just publish that once a month..."*
>
> *(Technical lead, IATI publishing government department)*

Whilst some aspects of making data commensurate with a standard can be fixed at the interface between internal systems and external data publication, others require changes to the source data itself. For example, the IATI standard invites particular categorisation, description and geolocation of aid activity information, which often requires additional labour from field staff and project officers to supply suitably detailed and structured information and to keep records up to date according to external open data publication schedules, instead of just according to internal management milestones.

Across these cases then we see that aligning with a standard transformed both the data published, and the data publisher: creating new and dynamic organizational structures and practices that did not previously exist. As with metadata standards (Bowker et al., 2010), GTFS, CSV and IATI all produce infrastructural changes in the organisations that adopt them changing technical activities and wider organizational arrangements.

Measuring Performance of Open Government Data Policies

Thus far, we have focussed on standards in terms of interoperability. However, the term standard is often also used in the context of performance standards: checking whether some phenomena measures up against some agreed minimum level of quality, or some criteria for success (Busch, 2011; Bruno & Didier, 2013). The open data standards we have explored have come to be used as means of operationalising assessments of whether OGD initiatives are delivering against principles of machine readability, or against specific transparency goals.

For instance, the Ile-de-France region, in an internal presentation made public to data producers, uses Tim Berners Lee's 5-star scale for Linked Open Data which sets criteria based on the use of open standards for assessing data quality. This internal document considers that the mere use of CSV, instead of the Excel XLS format for example, increases the quality of the dataset without even looking at its content. The same goes for the use of the GTFS format. One interviewee, a database manager in a transit agency explained that with GTFS *"it allows [you] now to have quality data"* (Database manager, transit organisation, France). The UK has gone further in treating a technical assessment of file format as a measure of 'openness', in 2012 introducing to data.gov.uk a feature, which translated the 5-star scale into an algorithm that grades every datasets on the portal. The team in charge of data.gov.uk justified the publicity of the scores, shown on each dataset page as an 'openness rating' and available to explore as an average per publishing agency, as a *"useful driver to improve the data." (Data.gov.uk, 2015)*

Besides being a sign of a quality dataset, open standards are also used to gauge the advancement of the open data program itself. When launching their scoring algorithm, the UK government announced that *"the average openness score for all departments is 52%, based on the percentage of the datasets published by each department and its arms-length bodies that achieve 3 stars""* (Gov.uk, 2012).). Here, the use of an open file format for publishing data is being used as a proxy for the openness of a government department. For the

departments which scored poorly, the Cabinet Office announced it will undertake measures to *"improve their performance"* most notably by producing stronger guidance on how to publish data. Yet, such measures don't look inside the dataset to see whether the data is well structured, accurate or meaningful: they simply assess the container.

Along similar lines in December 2011, after the launch of data.gouv.fr, Regards Citoyens, published a blog post in reaction to the new portal. Entitled "Open data: an average grade for a data.gouv.fr under proprietary formats", the article assesses the new open data portal by the standard in use after an examination of the catalogue: *"We were able to find only a dozen of CSV and XML datasets against several hundreds under Microsoft proprietary formats. A serious effort still needs to be accomplished on this matter. According to the norm of the inventor of the web, it is only a small average grade we can grant data.gouv.fr for its launch."* (Regards Citoyens, 2011) This example again shows again the importance being placed upon open formats as a sign of a "good" open data policy. Tim Berners Lee's 5-star rating has become much more than simple guidelines for opening data: publishing data in open formats regardless of the quality or the content of the data is taken to indicate that the open data program is conducted in a good direction.

In the case of the International Aid Transparency Initiative, the IATI standard has also become operationalized as a domain specific measurement tool. The advocacy organisation Publish What You Fund has created an Aid Transparency Index (ATI) every year since 2011. The ATI, originally based on a manual survey of data provision by government aid donors, now uses indicators *"selected using the information types agreed in the International Aid Transparency Initiative (IATI) standard"*, and weights scores on 22 indicators at least 50% lower if the data is available, but not structured using the IATI schema (Publish What You Fund, 2011). The ATI retains considerable components of manual data collection alongside automated assessment of IATI data, yet it points to the way in which standards can influence the way in which assessments of transparency may be carried out.

However, claims about what makes for 'good' open data performance are not neutral claims: they build in broad assumptions about who the primarily users of government data are, and how that data should be used.

Configuring Data Towards Advanced Users

As we have explored at the implementation of data standards in each of our cases, we have found tensions concerning how usable the standardised data is, and by whom. CSV, GTFS and IATI each place emphasis on a particular kind of machine-readability, and their use ultimately aims at reaching a certain type of user who has the capacity to achieve anticipated goals of the open data project. Indeed the main goal of many open data projects is framed in terms of encouraging reuse of published data to create websites, apps and services that in turn will generate economic, social and political value. This requires the data to reach users who have the technical skills to reuse the data, but also the potential capacity to create services without direct funding from public bodies. For the open data project managers we interviewed, there was a common identification of these users as professional developers inspired by the free/open source movement. As such, the use of open standards in the exchange of information was as much about a cultural practice that encourages the development of ecosystems around the published datasets (Russell, 2014), as it was a practical step to make the data easier to work with. When publishing a dataset, an open data project manager in a French city explained she had to make choices on the format she will use:

> *"I put the data in multiple formats to try to find a balanced choice between a very raw format, for example CSV which is something very usable for developers even if it is a bit less for people who just want to see what the data look like."*
>
> *(An open data project manager, French city)*

Here the use of CSV involves a practical choice that orientates the open data policy. The choice of CSV format will increase the frictions for using the datasets by the general public who might be con-

fused by the settings required to open the file. On the other hand, it may appeal to developers who can directly import the machine-readable dataset in the tools they use or create.

The same choice appears in the case of choosing the GTFS format. The complexity of using a GTFS file is far greater than CSV, as the dataset is divided in multiple related files. As a result it is mostly professional developers and transport specialists with the capacity to use data in this format. In a local transit agency, the choice of the format raised some concern that its complexity would reduce the user base to very skilled developers:

> *"We release the data in a format which is really well done, it's a Google norm but it can be very complex to understand. We told ourselves 'the guy who will be able to release an app with that, is going to be really solid'"*
>
> *(A database manager, local transit agency)*

But the choice of this standard was also in many cases driven by demand from developers:

> *"Our problem was that we asked ourselves 'but in format should we publish the data? GTFS?' We did not really know. […] What we did, we went ask the developers but of course we discussed this between technicians and the developers told us 'In our opinion, GTFS is a good format, popular, documented, easy to access, let's start with that'"*
>
> *(An open data project manager, local transit agency)*

This quote indicates that the project managers followed the recommendations of technically skilled developers in order to increase the usage of the data. However, had the project manager been in conversation with other potential data users, they potentially would have had other answers as to what would make the data more usable to them. Open data are thus often being calibrated to the expectations and needs of the users closest to the officials releasing data: these relationships in effect acting against the implicit idea that open data should be configured to be equally open to all.

In the case of IATI, there has been a long journey to bridge the needs of data providers and users. The early publishers of IATI data, and those involved in governing the design of the standard, were large government aid donors, with established ICT departments charged with generating their open data directly out of large Enterprise Resource Planning (ERP) systems or internal project databases. For these users, structured XML was familiar, and allowed flexibility in expressing their data. However, the community of users around aid information is much less technically adept – consisting in analysts who are much more comfortable in using spreadsheets of tabular data than they are coding to work with nested data structures. As an increasing number of Non Governmental Organisations, with much more limited technical capacity, have entered the community of data publishers, there have been substantial efforts towards the creation of tools and services that can work with IATI XML, either providing web interfaces that hide the underlying data formats entirely, or providing conversion tools that convert between IATI XML and CSV flat file formats. Reflecting on supporting publishers and users of IATI data, one member of community noted:

> *"When you start talking about XML and showing people what XML looks like, and [...] [how an] [...] XML file is different to a CSV file, why it's better to use XML rather than having lots of spreadsheets, they tend to start running for the door... But I think there are ways of explaining IATI and also publishing using something like Aidstream [an online publishing platform] where you don't have to even engage with [...] XML."*
>
> *(A member of the IATI support team)*

There is an explicit recognition around IATI of the need for intermediary platforms, which will sit between many users and the data. The same may be said to be true in the case of GTFS, where skilled users configure map-based interfaces or apps to allow others to access transport information. However, whilst GTFS intermediaries tend to be converting data into information, the limited number of people in the community around IATI with in-depth XML skills leads to a layer of inter-

mediaries converting data into data (Davies, 2010), reformatting from a shared structured standard, into a proliferation of non-standardised flattened file formats.

As open government data standards work to configure data towards advanced users, they also introduce other layers of infrastructure and intermediation between citizens and the state. Contrast the direct request to government for information under Right to Information based transparency, in which the citizen is able to demand an account direct from the state, with the mediated access to information presented through OGD. Our point here is ultimately descriptive, not normative: in some cases, the OGD approach may deliver greater effective transparency – but we have to be attentive to the role that the operationalization of machine-readability and through open data standards is playing in shaping who has data, and how they can access it.

Discussion and Conclusions

When looking at the back rooms of open data, the requests that governments 'use CSV', 'publish data as GTFS', or 'adopt the IATI standard for their data' involve much more than a simple operation in which data producers would use the 'save as' menu item and switch the format. Instead, at a variety of levels, standards are substantively shaping not only the production of open data, but are also leading to quiet and localised transformations of bureaucracies. We have seen how standards formats and schema are increasingly becoming indicators of the advancement of open data programs, and adoption of standards as part of open data publication is seen as a crucial part of enacting an open data agenda, realising core principles of making data machine-readable. In response, government officials are engaging in work processes to turn the spreadsheets used on the desktops of their colleagues, and the internal databases from specific departments, into standardised datasets optimised for a particular kind of machine-readability outside of the state: constructing their 'raw' datasets in the process. Mediated via standards, the transparency delivered by OGD reveals one particular rationalisation and representation of the information held inside the state, focussing on machine-mediated transparency, rather than transparency as a relationship between citizen, and account-giving state. The particular affordances of open data formats and standards, with their emphasis on machine-readability, act as a filter on what can or can't be easily expressed as part of OGD transparency.

However, we must also recognise that the histories embedded in the formats and standards being adopted owe as much to politics as they do to technology. Though often appearing as recent creations, open data standards are embedded in much deeper information infrastructures (Star, 1999). Each format and standard we have explored builds on legacy practices going back decades, whether at the level of formatting, as for XML and CSV, or at the level of the categories and classifications built into the standard, as in the case of IATI, and it's reliance on terms derived from OECD political systems, and its data structure defined through a process of political negotiation. These histories are inscribed into each of the datasets created using the standards, although few data publishers or users may be consciously aware of them at the point of publication or use. For GTFS, for example, the fact that it can express timetables, but not public transport performance, for example, is rarely considered when it is selected as *the* format for publishing transport information (Rojas, 2012). Crucially then, and counter to the tone of much open data discourse, the 'openness' of open data does not mean that it is freed from past politics, or from previous generations of technology, which, through their role in defining the information infrastructures from which data is drawn and the standards via which its open incarnation is represented, continue to influence what gets expressed: what is made visible, and what, in effect, disappears when moving data from inside the organisation to the open data domain.

Unlike the negotiated metadata standards for scientific data sharing, shaped within relatively defined communities of practice, the open government data standards we have explored are generally experienced by the data workers of the state as fixed points. Data workers are tasked to implement the standard, and organise their work practices accordingly, but they have limited practical capability to shape the standards around

their local needs. Indeed, incorporating user needs in the dataset is often experienced as problematic by data producers and standard-setters (Denis & Pontille, 2013). Divergent user requirements risk disruption to the abstract and idealised application of the standards, and threaten the goal of 'frictionless', globally interoperable data. Data standards can thus come to stand in place of dialogue with the community of potential users of data. Of course, it is not that these standards have no notion of the user. Choices made over the standards in our case studies configure the data (Woolgar, 1991) to advanced users with the skills to open and reuse them. The materiality of machine-readable datasets anticipates the skills and materializes a certain representation of the user (Akrich, 1992). But the standards in use also create multiple data publics (Ruppert, 2012): developers who can reuse the data to create services, advanced users who can open the dataset and do basic analysis and the general public who are expected to benefit from the opening of data only via intermediaries, using what Ruppert (2012) calls "literary technologies" such as visualisations, maps, applications and online services to gain second-hand access to the disclosures made by the state. By making data machine-readable, open data standards, in theory, allow machines to join the cohort of "armchair auditors" (Ruppert, 2012) producing a particular notion of accountability in the transparency agenda of open data policies.

Open government data standards also exist in a context that tightly couples the technical, social, and organisational with the explicitly political: policy commitments have been publicly made against open data principles. Yet, because open data is framed as being published for anyone to re-use, the test of successful publication cannot be any one specific use of the data, but has to instead be a proxy for potential usability of the data. Evaluation of the standardisation and machine-readability of data has increasingly become this proxy within OGD policy making. Yet, setting format standards as the metric for a good open data initiative can leave the content of a dataset entirely out of the picture. Although a focus on schema standards brings in greater consideration of what the data contains, it still stands in for any evaluation of OGD against ultimate goals of creating transparency, in which evaluations would need to address not only the broadcast of data, but also it's effective receipt and re-use.

The account we have offered in this paper provides an initial overview of just three different standards, operating at different levels within the growing landscape of open data. In drawing on empirical work to present a descriptive account of these standards in action at the point of data production, we have sought to contribute an open data component towards the called for development of 'critical data studies' (Kitchin & Lauriault, 2014), unpacking these data standard assemblages, and looking at their materiality within the context of public organisations. This is by no means an argument against adoption of standards: rather, it is an account intended to support constructive and critical approaches to their evaluation and adoption. Further work is needed to trace forward the consequences of these data standards assemblages, and current orientations towards the machine-readability of data, in producing new transparency regimes of the state. In these relatively early days of open government data standardisation, with a new layer of 'open data infrastructure' being built out through the work of policy-makers, technologists and data workers, developing these approaches to bring standards, their stories, and their possible consequences, into view, requires ongoing attention.

References

Akrich M (1992) The De-scription of Technical Objects. In: Bijker W & Law J (eds) *Shaping Technology—Building Society: Studies in Sociotechnical Change*. Cambridge, MA: The MIT Press.

Almklov P G (2008) Standardized Data and Singular Situations. *Social Studies of Science* 38(6): 873-897.

Atz U, Heath T, & Fawcett J (2015) *Benchmarking Open Data Automatically* (No. ODI-TR-2015-000). London: The Open Data Institute.

Bates J (2012) "This is what modern deregulation looks like": co-optation and contestation in the shaping of the UK's Open Government Data Initiative. *Journal of Community Informatics* 8(2).

Bates J (2013) The Domestication of Open Government Data Advocacy in the United Kingdom: A Neo-Gramscian Analysis. *Policy and Internet* 5(1): 118-137.

Berners-Lee T (2009) "The Next Web". TED talk. Available at: http://www.ted.com/talks/tim_berners_lee_on_the_next_web (accessed: 10.11.2016).

Berners-Lee T (2010) Design issues – Linked Data. Available at: http://www.w3.org/DesignIssues/LinkedData.html (accessed: 10.11.2016).

Birchall C (2014) Radical Transparency? *Cultural Studies Critical Methodologies* 14(1): 77-88.

Borgman C L (2007) *Scholarship in the digital age: Information, infrastructure and the internet*. Cambridge: The MIT Press.

Bowker G C (1994) *Science on the Run: Information management and industrial geophysics at Schlumberger, 1920-1940*. Boston, MA: The MIT Press.

Bowker GC (2000). Biodiversity datadiversity. *Social Studies of Science* 30/5:643-683.

Bowker et al. (2010) Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment. In: Hunsinger et al. (eds) *International Handbook of Internet Research*. New York:Springer, 97–117.

Bowker G C & Star S L (2000) *Sorting Things Out: Classification & its Consequences.* Cambridge, MA: The MIT Press.

Busch L (2011) *Standards: recipes for realities*. Cambridge, MA: The MIT Press.

Bruno I & Didier E (2013) *Benchmarking :* L'état sous pression statistique. Paris: Zones.

Cabinet Office (2013) National Information Infrastructure. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/254166/20131029-NII-Narrative-FINAL.pdf (accessed: 10.11.2016).

Chignard S (2013) A brief history of open data. Paris Tech Review. Available at: http://www.paristechreview.com/2013/03/29/brief-history-open-data/ (accessed: 10.11.2016).

Davies T (2012) Emerging Implications of Open and Linked Data for Knowledge Sharing in Development. *Institute of Development Studies Bulletin* 43(5).

Davies T (2013) "There's no such thing as raw data". In: Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13, New York, USA: ACM Press, 75-78.

Davies T (2014) Five critical questions for constructing data standards. Available at: http://www.timdavies.org.uk/2014/02/21/five-critical-questions-for-constructing-data-standards/ (accessed: 10.11.2016).

DATA Act (2014) S.994 - 113th United States Congress (2013-2014): Digital Accountability and Transparency Act.

Data.gov.uk (2015) Five Stars of Openness. Available at: http://guidance.data.gov.uk/five_stars_of_openness.html (accessed: 10.11.2016).

DCLG (2014) Local Government Transparency Code 2014.

Denis J, & Pontille D (2012) Travailleurs de l'écrit, matières de l'information. Revue D'anthropologie Des Connaissances, 6(1):1-20.

Denis J & Pontille D (2013) Parasite Users? Users and Cycling Infrastructures in OpenStreetMap.

Denis J & Pontille D (2014) Parasite Users ? Users and Cycling Infrastructures in OpenStreetMap. In Mongili A & Pellegrino G (eds) *Information Infrastructure(s): Boundaries, Ecologies, Multiplicity.* Cambrige: Cambridge Scholars Publishing, 146-167.

Edwards P (2010) *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: The MIT Press.

Edwards P, Mayernik M S, Batcheller A L, Bowker G C, & Borgman C L (2011) Science friction: Data, metadata, and collaboration. *Social Studies of Science* 41(5): 667-690.

Espeland W & Stevens M (1998) Commensuration as a social process. *Annual Review of Sociology* 24(1998): 313-343.

Fecher B & Friesike S (2014) Open science: One term, five schools of thought. In: Friesike S & Bartling S (eds) *Opening science*. New York: Springer.

Flynn P (2014) The XML FAQ. Available at: http://xml.silmaril.ie/ (accessed: 10.11.2016).

Fumega S & Scrollini F (2011) Access to Information and Open Government Data in Latin America. Available at: https://spaa.newark.rutgers.edu/sites/default/files/files/Transparency_Research_Conference/Papers/Fumega_Silvana.pdf (accessed: 10.11.2016).

G8 (2013) G8 Open Data Charter. Available at: https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex (accessed: 10.11.2016).

Goëta S (2014) The Daily Shaping of State Transparency: Emerging Standards in Open Government Data. Presentation at the ESOCITE/4S Joint Meeting in Buenos Aires, August 20-23 2014.

Gov.uk (2012) New funding to accelerate benefits of open data. Available at: https://www.gov.uk/government/news/new-funding-to-accelerate-benefits-of-open-data (accessed: 10.11.2016).

Gurstein M (2011) Open data: Empowering the empowered or effective data use for everyone? *First Monday* 16(2).

Gurstein M (2012) Two Worlds of Open Government Data: Getting the Lowdown on Public Toilets in Chennai and Other Matters. *Journal of Community Informatics* 8(2).

Heald D (2006) Varieties of Transparency. *Proceedings of the British Academy* 135: 25-43.

Hüsemann S (2001) Information Exchange Platform for Humanitarian Development Projects. Available at: https://www.researchgate.net/publication/237243797_Information_Exchange_Platform_for_Humanitarian_Development_Projects (accessed: 10.11.2016).

Janssen M, Charalabidis Y & Zuiderwijk A (2012) Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management* 29(4): 258-268.

International Open Data Charter (2015) Accessible at: http://opendatacharter.net (accessed: 10.11.2016).

Johnson JA (2013) From open data to information justice. In *Annual conference of the* midwest political science association, Chicago (IL), April 13 2013.

Karasti H, Millerand F, Hine C M, Bowker G C (2016) Guest Editorial: Knowledge infrastructures: Part 1. *Science & Technology Studies* 29(1): 2-12.

Khan S & Foti J (2015) Aligning supply and demand for better governance: open data in the open government partnership. Available at: http://www.opendataresearch.org/dl/symposium2015/odrs2015-paper49.pdf (accessed: 10.11.2016).

Kitchin R & Lauriault T P (2014) Towards critical data studies: Charting and unpacking data assemblages and their work. In: Eckert J, Shears A & Thatcher J (eds) *Geoweb and Big Data*. University of Nebraska Press.

Kitchin R (2014) *The data revolution: Big data, open data, data infrastructures & their consequences*. London: Sage Publications.

Lampland M & Star S L (2008) *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*. Itahaca: Cornell University Press.

Larsson T (1998) How open can a government be? The Swedish experience. In Deckmyn V & Thomson I (eds) *Openness and Transpare*ncy. Maastricht: European Institute of Public Administration.

Lathrop D & Ruma L (2010) *Open government: Collaboration, Transparency, and Participation in Practice*. Cambridge, MA: O'Reilly.

Law J & Mol A (2008) The Actor-Enacted: Cumbrian Sheep in 2001. In: Malafouris L & Knappett C (eds) *Material Agency*. Boston, MA: Springer US, 57-77.

Levien R (1998) The decommoditization of protocols. Available at: http://www.levien.com/free/decommoditizing.html (accessed: 10.11.2016).

Local Government Association (LGA) (2015) Local transparency guidance – publishing data. Available at: http://www.local.gov.uk/practitioners-guides-to-publishing-data (accessed: 10.11.2016).

Millerand F & Bowker G C (2009) Metadata standards: Trajectories and enactment in the life of an ontology. In: Lampland M & Star S L (eds) *Standards and their stories*. Itahaca: Cornell University Press, 149-165.

Morozov E (2013) *To Save Everything, Click Here: To Save Everything, Click Here: The Folly of Technological Solutionism*. Scientific American. New York, USA: Public Affairs.

Murillo MJ (2014) Evaluating the role of online data availability: The case of economic and institutional transparency in sixteen Latin American nations. *International Political Science Review* 36: 42–59.

Obama B (2009) Memorandum on Transparency and Open Government. Available at: http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment (accessed: 10.11.2016).

Obama B (2013) Executive Order - Making Open and Machine Readable the New Default for Government Information. Available at: https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-(accessed: 10.11.2016).

Open Data Institute (2014) 2014: The Year of CSV. Available at: https://theodi.org/blog/2014-the-year-of-csv (accessed: 10.11.2016).

Open Knowledge Foundation (OKF) (2011) Beyond Access: Open Government Data & the Right to (Re) use Public Information. Available at:: http://www.access-info.org/documents/Access_Docs/Advancing/Beyond_Access_7_January_2011_web.pdf (accessed: 10.11.2016).

Open Stand (2012) OpenStand: Principles for The Modern Standard Paradigm. Available at: http://openstand.org/ (accessed: 18.10.2014).

Palme J & Pargman D (2009) ASCII Imperialism. In: Lampland M & Star S L (eds) *Standards and their stories*. Itahaca: Cornell University Press.

Peixoto T (2013) The uncertain relationship between open data and accountability: A response to Yu and Robinson's 'the new ambiguity of open government'. *UCLA Law Review* 60(200): 200-213.

Pollock R (2007) Give Us the Data Raw, and Give it to Us Now. Open Knowledge's Blog. Available at: http://blog.okfn.org/2007/11/07/give-us-the-data-raw-and-give-it-to-us-now/ (accessed: 14.11.2016).

Pollock R(2013) Frictionless Data: making it radically easier to get stuff done with data. Blog post in Open Knowledge International blog. Available at: http://blog.okfn.org/2013/04/24/frictionless-data-making-it-radically-easier-to-get-stuff-done-with-data/#sthash.tsCuAV3h.dpuf (accessed: 10.11.2016).

Publish What You Fund (PWYF) (2011) Aid Transparency Index: Methodology and Data Sources. Available at: http://www.publishwhatyoufund.org/index/ (accessed: 10.11.2016).

Regards Citoyens (2012) OpenData: La moyenne pour un data.gouv.fr sous formats propriétaires. Available at: https://www.regardscitoyens.org/opendata-la-moyenne-pour-un-data-gouv-fr-sous-formats-proprietaires/ (accessed: 10.11.2016).

Robinson D G, Yu H, Zeller W, & Felten E (2009) Government data and the invisible hand. *Yale Journal of Law & Technology* 11(160): 160-175. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1138083 (accessed: 10.11.2016).

Rojas F M (2012) Transit Transparency. Available at: http://www.transparencypolicy.net/assets/FINAL_UTC_TransitTransparency_8%2028%202012.pdf (accessed: 10.11.2016).

Ruppert E (2012) Doing the Transparent State: open government data as performance indicators. In: Mugler J & Park S-J (eds) *A World of Indicators: The production of knowledge and justice in an interconnected world*. Cambridge: Cambridge University Press, 51-78.

Russell A (2014) *Open Standards and the Digital Age: History, Ideology, and Networks.* Cambridge: Cambridge University Press.

Star SL (1999) The Ethnography of Infrastructure. *American Behavioral Scientist* 43(3): 377-391.

Star S L & Ruhleder K (1996) Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research* 7(1): 111-134.

Strasser B J (2012) Data-driven sciences: From wonder cabinets to electronic databases. *Studies in History and Philosophy of Biological and Biomedical Sciences* 43(1): 85-7.

Suchman L (1995) Making work visible. *Communications of the ACM* 38(9): 56-64.

Sunlight Foundation (2010) Ten Principles for Opening Up Government Information. Available at: https://sunlightfoundation.com/policy/documents/ten-open-data-principles/ (accessed: 10.11.2016).

Woolgar S (1991) Configuring the User. In: *A Sociology of Monsters: Essays on Power, Technology and Domination*. London: Routledge, 57-103.

W3C (2013) CSV on the Web Working Group Charter. Available at: http://www.w3.org/2013/05/lcsv-charter.html (accessed: 10.11.2016).

W3C (2015) Data on the Web Best Practices: W3C First Public Working Draft 24 February 2015. Available at:: http://www.w3.org/TR/2015/WD-dwbp-20150224/ (accessed: 10.11.2016).

Web Foundation (2015) Open Data Barometer: Second Edition. Available at: http://opendatabarometer.org/assets/downloads/Open%20Data%20Barometer%20-%20Global%20Report%20-%202nd%20Edition%20-%20PRINT.pdf (accessed: 10.11.2016).

Weinstein J & Goldstein J (2012) The Benefits of a Big Tent: Opening Up Government in Developing Countries. *UCLA Law Review Discourse* 38(2012): 38-48.

Yu H & Robinson D G (2012) The New Ambiguity of "Open Government." *UCLA Law Review* 178(2012): 178-208.

Zimmerman A (2007) Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries* 7(1-2): 5-16.

Zimmerman A S (2008) New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. *Science, Technology & Human Values* 33(5): 631-652.

Zuiderwijk A & Janssen M (2014) Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly* 31(1): 17-29.

Zuiderwijk A, Janssen M, Choenni S, Meijer R, & Alibaks R S (2012) Socio-technical impediments of open data. *Electronic Journal of E-Government* 10(2): 156-172. Available at: http://www.ejeg.com/issue/download.html?idArticle=255 (accessed: 10.11.2016).

Notes

1 www.5stardata.info (accessed: 14.11.2016)
2 https://csvlint.io (accessed: 14.11.2016)
3 http://fr.slideshare.net/christophelibertidf/bonnes-pratiquesexcel-cc27juin2013 (accessed: 14.11.2016)