

INTRODUCTION

The dataset under consideration for deposit in our Open Access Repository is quite small (1.5 MBs) and will not require significant resources to store and preserve. As I will discuss at greater length below, the primary investigator has provided us with eight comma separated value files, one Excel file, and one pre-print version of a published research article in .pdf form. The data are a collection of aggregated tweets (or microblogs) from the social media network Twitter. They are all related to the 2013 national election in Italy, and collected by faculty (Fabio Giglietto and Donatella Selva) in the Department of Communications at the University of Potato. The data are used by the researchers to consider the new sociopolitical ecology of social media and the pressure it puts on traditional media. In this instance, they look at peak twitter activity and key terms (or “hashtags”) during specific talk shows over the course of the election season. As they note in their abstract, this is the first study of a complete dataset of Tweets (2,489,669) that span an entire season of a TV genre (1,077 talk-show episodes). A content analysis of the Tweets created during the 2013 Italian election season depicts different forms of participation (both by audience and political actors) and describes the roles new media and traditional media played at the time.

First, however, before we go over the technical and curatorial details, I want us to consider two policy issues regarding the nature of the data and why we should include it—and data like it—in our repository.

CURATING EMBEDDED DATA

This issue can be put better in the form of a question: Why are we considering the preservation and maintenance of digital assets that could be deposited in data repositories such as Dryad or Figshare? The short answer, of course, is that our mission includes preservation of the kinds of assets. Part of this mission is preserving data as embedded, available and accessible along with published material and articles. Although a pre-publication version of the researchers’ journal article is on the Social Science Research Network (SSRN) and it has been peer-reviewed and published by the *Journal of Communication*, neither of these entities will embed the data. Even if our depositor gains the advantages of exposure through SSRN and professional prestige from publication in peer-reviewed journal, our repository can preserve this data in its intended context for future reuse and methodological modeling.

CURATING SOCIAL MEDIA DATA

What makes this data particularly unique, however, is its source: *social media*. It represents an exceptional conjunction between social science data (political metrics) and humanities research (communications studies). In this instance, the data is not only measuring facts, it also has the potential to help in the interpretation of social sentiment, human behaviors, social psychology, and even emotions. These are areas that have traditionally brought together the social sciences and the humanities. The research is also concerned with the effectiveness of communication and rhetoric, which too have long been a key component in the humanities.

Burgess and Bruns claim that with the “Twitter API (the Application Programming Interface, which provides structured access to communication data in standardized formats) it is possible, with a little effort and sufficient technical resources, for researchers to gather very large archives of public tweets concerned with a particular topic, theme or event. Essentially, the API delivers very long lists of hundreds, thousands, or millions of tweets, and metadata about those tweets; such data can then be sliced, diced and visualized in a wide range of ways, in order to understand the dynamics of social media communication. Such research is frequently oriented around pre-existing research questions, but is typically conducted at unprecedented scale, [and] broadly concerned with understanding the role of social media in the contemporary media ecology, with a focus on the formation and dynamics of interest- and issues-based publics.”¹

Moreover, in the digital humanities too often we are limited from allowing such data to be open to use, but in this instance the data is both anonymized and not subject to copyright. We also have the opportunity to curate, rather than simply store or containerize, data that will have a good chance of being reused in future research and providing a model and methodology for other work.

DATA CURATION PROFILE: INFORMATION WE HAVE

Destination: Open Access Repository

Researchers' Information: Giglietto & Selva

Collection Title: Twitter Data Italian Election 2013

Is this appropriate material for our repository? **Yes**

- Material is a relatively new form of obtaining and aggregating social media data.
- Social media data from the immediate post-Berlusconi era in Italy will have long-term value.
- Serves our mission of open data access.

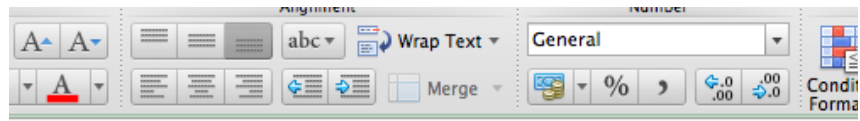
Is there confidential information or human subject material in the dataset? **No**

- Data is in aggregate form
- No individuals or identities are disclosed

Is the dataset controlled by any form of intellectual property? **No**

- The dataset files are not covered under copyright
- CSV files do not constitute a rights-controlled database

For context, note how the following image of some .xls columns in the dataset do not disclose individual identities and Twitter handles. Each column represents an aggregation of sometimes hundreds of thousands of users.



| | span | audience | share | id | tweet | contributors | reach |
|-----|------|------------|-------|--------------|-------|--------------|----------|
| lay | 170 | 781000 | 4.67 | piazzapulita | 3445 | 1003 | 5469654 |
| urs | 210 | 1017855 | 4.26 | piazzapulita | 7536 | 2035 | 8235104 |
| on | 130 | 209850 | 4.53 | omnibus | 37 | 13 | 11316 |
| on | 60 | 1390425 | 5.34 | ottoemezzo | 1118 | 519 | 1242395 |
| on | 170 | 674405 | 3.5 | infedele | 1037 | 410 | 702662 |
| es | 130 | 215000 | 4.47 | omnibus | 40 | 33 | 18283 |
| es | 60 | 1312000 | 8 | ottoemezzo | 554 | 285 | 396782 |
| es | 165 | 3102000 | 12.79 | ballaro | 5803 | 2136 | 4994400 |
| es | 130 | 870000 | 9.2 | portaaporta | 477 | 282 | 557185 |
| ed | 130 | 127000 | 3.05 | omnibus | 17 | 8 | 5758 |
| ed | 60 | 1596074 | 6.09 | ottoemezzo | 1354 | 474 | 3393555 |
| ed | 130 | 873000 | 9.91 | portaaporta | 157 | 75 | 49894 |
| urs | 130 | 200397.844 | 4.06 | omnibus | 16 | 14 | 31068 |
| urs | 60 | 1447810 | 5.42 | ottoemezzo | 319 | 180 | 187742 |
| urs | 170 | 1151199 | 5.26 | piazzapulita | 8006 | 2369 | 14005115 |
| urs | 130 | 1197000 | 13.56 | portaaporta | 65 | 52 | 23710 |
| l | 130 | 248000 | 5.08 | omnibus | 34 | 26 | 11466 |
| i | 60 | 1049056 | 4.14 | ottoemezzo | 406 | 192 | 299285 |
| i | 120 | 620000 | 8.52 | ultimaparola | 2829 | 472 | 2518135 |
| t | 130 | 207000 | 4.14 | omnibus | 70 | 38 | 51048 |
| t | 150 | 776000 | 3.46 | inonda | 1447 | 463 | 1011398 |
| n | 130 | 178000 | 3.59 | omnibus | 27 | 10 | 6532 |
| n | 70 | 804000 | 3.33 | inonda | 753 | 337 | 623644 |
| on | 130 | 288750.589 | 5.85 | omnibus | 46 | 25 | 20721 |
| on | 60 | 1567000 | 5.68 | ottoemezzo | 520 | 252 | 265656 |
| on | 170 | 927882 | 4.27 | infedele | 1951 | 863 | 1525440 |
| on | 130 | 1141000 | 12.73 | portaaporta | 1082 | 497 | 759661 |
| es | 130 | 193487.574 | 3.92 | omnibus | 73 | 48 | 45763 |
| es | 60 | 1720487 | 6.14 | ottoemezzo | 1040 | 504 | 865901 |
| es | 165 | 3583000 | 13.82 | ballaro | 7103 | 2212 | 6408703 |
| es | 130 | 828000 | 11.22 | portaaporta | 41 | 33 | 41041 |
| ed | 130 | 155000 | 3.52 | omnibus | 21 | 18 | 15033 |
| ed | 60 | 1312103 | 4.64 | ottoemezzo | 529 | 245 | 273014 |

For comparison, a common method of using the Twitter API or a public web data scrape includes a significant amount of user identification. This particular screen shot is from Giglietto's own website (larica.uniurb.it/nextmedia) and collates tweets using the hashtag (#ir14) for a recent academic conference.

```

Judging by #ir14 tweets, its the
best internet conference ever.</data>
<data key="V-Layout Order">1.55433630035443</data>
<data key="V-X">5625.8271484375</data>
<data key="V-Y">7610.1318359375</data>
<data key="V-In-Degree">2</data>
<data key="V-Out-Degree">11</data>
<data key="V-Betweenness Centrality">1811.680205</data>
<data key="V-Closeness Centrality">0.000464</data>
<data key="V-Eigenvector Centrality">0.00266</data>
<data key="V-PageRank">1.40635</data>
<data key="V-Clustering Coefficient">0.254545454545455</data>
<data key="V-Reciprocated Vertex Pair Ratio">0</data>
<data key="V-ID">4</data>
<data key="V-Followed">710</data>
<data key="V-Followers">2156</data>
<data key="V-Tweets">57210</data>
<data key="V-Favorites">15285</data>
<data key="V-Time Zone UTC Offset (Seconds)">3600</data>
<data key="V-Description">My mind is digital, my heart in the southern bit of Africa, me
http://t.co/L21DcFA9hE , doc maker</data>
<data key="V-Location">Hackney, London</data>
<data key="V-Web">http://t.co/N6hMegeHPq</data>
<data key="V-Time Zone">London</data>
<data key="V-Joined Twitter Date (UTC)">11/12/2007 7:17:42 PM</data>
<data key="V-Custom Menu Item Text">Open Twitter Page for This Person</data>
<data key="V-Custom Menu Item Action">http://twitter.com/wildebees</data>
<data key="V-Tweeted Search Term?">Yes</data>
<data key="V-Top URLs in Tweet by
Count">https://www.conftool.com/aoir-ir14/sessions.php?utm_source=buffer&utm_campaign=Buff
ter</data>
<data key="V-Top URLs in Tweet by
Salience">https://www.conftool.com/aoir-ir14/sessions.php?utm_source=buffer&utm_campaign=B
witter</data>
<data key="V-Top Domains in Tweet by Count">conftool.com</data>
<data key="V-Top Domains in Tweet by Salience">conftool.com</data>
<data key="V-Top Hashtags in Tweet by Count">ir14 IR14</data>
<data key="V-Top Hashtags in Tweet by Salience">IR14 ir14</data>
<data key="V-Top Words in Tweet by Count">ir14 data w nancybaym call researchers know al
<data key="V-Top Words in Tweet by Salience">w data nancybaym call researchers know algo
<data key="V-Top Word Pairs in Tweet by Count">rt,nancybaym nancybaym,call call,resear
algorithms,use use,apps apps,visualize visualize,data data,jeffhemsley</data>
<data key="V-Top Word Pairs in Tweet by Salience">rt,nancybaym nancybaym,call call,res
algorithms,use use,apps apps,visualize visualize,data data,jeffhemsley</data>
</node>
<node id="wikiresearch">
<data key="V-Shape">Image</data>
<data key="V-Size">162.557178593854</data>
<data key="V-Opacity">99.9989870280702</data>
<data key="V-Image File">http://pbs.twimg.com/profile_images/1901640487/W normal png</da

```

In regard to privacy and human subject issues, standards for anonymizing so-called “public” social media data are still being developed. Indeed, there may be an urgent need for such standards. dana boyd reminds us that “Wanting privacy is not about needing something to hide. It’s about wanting to maintain control. Often, privacy isn’t about hiding; it’s about creating space to open up. If you remember that privacy is about maintaining a sense of control, you can understand why Privacy is Not Dead. . . . [All of us] need to think through the implications and ethics of our decisions, of what it means to invade someone’s privacy, or how our presumptions about someone’s publicity may actually affect them.”²

The researchers in this instance successfully used aggregation techniques to collect and mine their data without including individual identities. In the future, it may become standard or even codified into law that data research using public social media identities becomes proscribed.

DATA CURATION PROFILE: INFORMATION WE NEED

Before proceeding with preparation of the data files, the depositors need to provide us with further information.

1. The **intellectual property** status of the research paper derived from the data is unclear. Did the authors amend any restrictive publication licenses to allow our repository to make the paper available on an open access basis?

- Pre-publication paper is deposited with Social Science Research Network (SSRN). Although “SSRN does not take a copyright for any papers (or other documents) on SSRN,” we do not know what licenses were made for the paper’s pre-publication status. doi: dx.doi.org/10.2139/ssrn.2345240³
- A revised version of the paper, “Second Screen and Participation: A Content Analysis on a Full Season Dataset of Tweets,” is also published in the *Journal of Communication* 64.2 (April 2014): 260–277. © 2014 International Communication Association and Wiley Online. doi: [10.1111/jcom.12085](https://doi.org/10.1111/jcom.12085)⁴

2. The data files provided to us lack even the most rudimentary **metadata**. Given the lack of metadata, we should consider using the CSV Validator developed by the British National Archives as a vehicle for obtaining metadata from them. digital-preservation.github.io/csv-validator/#toc8

- The Validator allows depositors to create metadata in a spreadsheet format, compensating for the lack of technical knowledge in either XML or RDF and the lack of experience of tools for producing and validating XML or RDF.
- Most users already have Microsoft Excel (or an equivalent) installed which they know how to use to produce a CSV file.

LICENSING

This dataset will be made available under the Open Database License (ODbL 1.0) according to the wishes of the depositors and in accordance with our own open access policies. The one exclusion may be the published research paper itself, which we hope to include. If permission is refused, I would still advise moving forward with the deposit of the dataset.

Any rights in the individual contents of the dataset are licensed under the Database Contents License. Users are therefore free:

- To Share: To copy, distribute and use the database.
- To Create: To produce works from the database.
- To Adapt: To modify, transform and build upon the database.

As long as they:

- **Attribute:** You must attribute any public use of the database, or works produced from the database, in the manner specified in the ODbL. For any use or redistribution of the database, or works produced from it, you must make clear to others the license of the database and keep intact any notices on the original database.
- **Share-Alike:** If you publicly use any adapted version of this database, or works produced from an adapted database, you must also offer that adapted database under the ODbL.
- **Keep open:** If you redistribute the database, or an adapted version of it, then you may use technological measures that restrict the work (such as DRM) as long as you also redistribute a version without such measures.

FILE INVENTORY

Total size of dataset: 1.5 MBs

| File name | File size | Short description |
|--------------------|-----------|-----------------------------------|
| peaks.csv | 44 KB | aggregation by show and hashtag |
| shows.csv | 2 KB | aggregation by show and tweets |
| episodes.csv | 138 KB | aggregation by show and episodes |
| Table_S1.xls | 70 KB | codebook examples |
| Table_S2.csv | 3 KB | typologies of scenes during peaks |
| Table_S3.csv | 31 KB | random sample of peaks |
| Table_S4.csv | 151 bytes | codebook |
| Table_S5.csv | 1 KB | frequency of typologies |
| Table_S6.csv | 1 KB | frequency of subgenres |
| SSRN-id2345240.pdf | 345 KB | pre-publication paper |

DATA INTEGRITY

The .csv format is not a particularly complicated format. There is little need for a tool such as JHOVE2. But that also doesn't mean that the integrity of the dataset—identifying null values, for example—isn't still necessary. Such a data integrity check should be run before we ask the depositors for further information about the dataset. The check may reveal more questions for them. Because .csv is used as a way to upload and download data in SQL data management systems, developers have written scripts on their own to check the integrity of .csv files. A recent example is Richard Pearce's **CSV Data Integrity - Table Analyser**, a QVW tool that does the following (source: community.qlik.com/docs/DOC-3352):

Date Check: The code checks each column where the name includes the word “Date”. This can be modified in the code. Each field value is checked against possible date formats and identified as either a good or bad date. Bad date distinct values are recorded for further analysis

Value Frequency: The code loads each distinct value for each column and counts the frequency they're encountered.

Distinct Values: Counts the number of distinct values for each field, again after converting the value to lower case and again after trimming leading & trailing spaces from the value.

Distinct N/A: Uses a mapping table to convert possible N/A entries into a consistent value and counts the number found for each column.

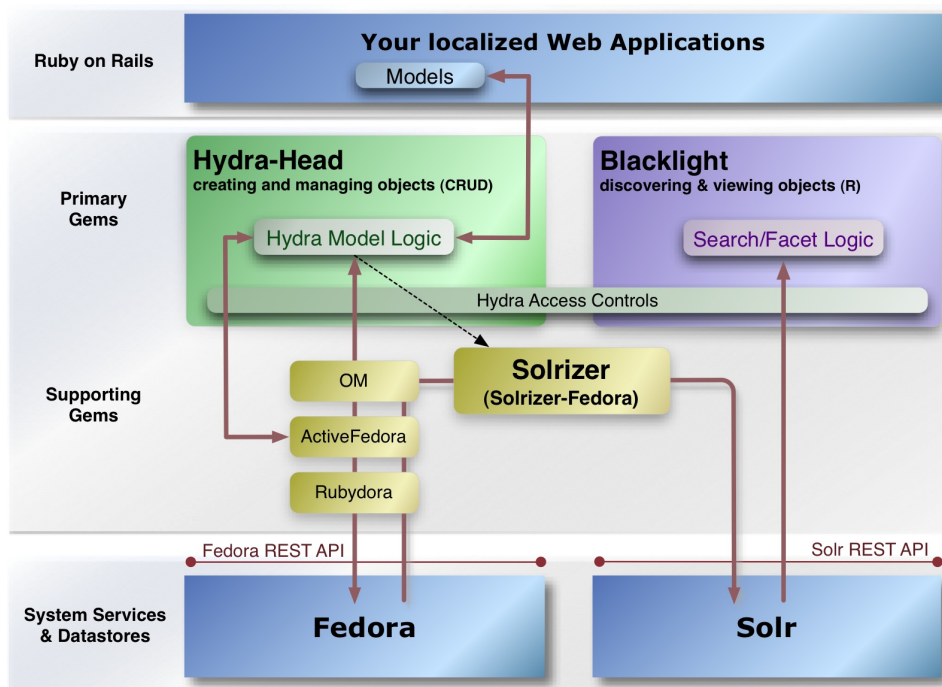
Null Values: Counts the number of empty records for each column

Data Types: Checks to see which data type is found for each value of a column. This could be either a number or text. If all fields are returned as Number or Text the field is assigned this property. If a mixture is returned this is recorded.

Such a tool should be utilized before moving forward with our own work on the dataset.

FORMAT INVENTORY: OAIS (SIP, AIP, DIP)

Our Fedora/Hydra stack should be operational by September 2014. We do not anticipate being ready to ingest this set of digital materials until after the stack is in place. The Hydra system will allow easier automation of necessary Fedora metadata and datastreams. Our stack will look something like the following figure:



SOURCE: WIKI.DURASPACE.ORG/DOWNLOAD/ATTACHMENTS/22022608/HYDRA+ARCHITECTURE+DIAGRAMV2.JPG

Initially, all files in this Twitter dataset will be converted to separate XML objects containing appropriate metadata schemas and unique identifiers. In terms of the OAIS Lifecycle curation model, the following protocols will be used.

1. **SIP:** We will ingest using the REST interface with Fedora Object XML (FOXML). Some reasons for this format are as follows (source: fedora-commons.org):

- FOXML as the internal storage format for Fedora objects enables easier evolution of functionality in the Fedora repository, without requiring ongoing extensions to other community formats.
- A PID is a unique, persistent identifier for a Fedora digital object. PIDs will be automatically assigned by a repository using the DOI standard.
www.doi.org/factsheets/DOIIdentifiers.html
- The URI for a Fedora object is constructed simply by appending the PID to the string "info:fedora/".
- FEDORA OBJECT PROPERTIES are non-versionable properties of the digital object. At ingest, the object **MUST** have these attributes: type:(required). The object is classified as one of three primitive Fedora object types, namely FedoraObject, FedoraBDefObject, or FedoraBMechObject state:(required) The object state can be Active (A), Inactive (I), or Deleted (D).

- And at ingest, it may have these optional attributes: label: (optional) The object is given a user-defined descriptive label; contentModel: (optional) The object is given user-defined content model identifier (to classify the pattern of datastreams and disseminations found in this object). The system will automatically assign these attributes (they should not be put in the ingest files, but they appear in stored files); createDate: (system assigned) The object creation date is assigned to the millisecond; lastModifiedDate:(system assigned) The object creation date is assigned to the millisecond.

2. **AIP:** Fedora Repository in FOXML format (source: wiki.duraspace.org) is designed, as well, for management and preservation:

- FOXML as the internal storage format for Fedora objects enables easier evolution of functionality in the Fedora repository, without requiring ongoing extensions to other community formats.
- The creation of Fedora digital object relationship metadata is the basis for enabling advanced access and management functionality driven from metadata that is managed within the repository. Examples of the uses of relationship metadata include:
 - Organize objects into collections to support management, OAI harvesting, and user search/browse
 - Define bibliographic relationships among objects such as those defined in Functional Requirements for Bibliographic Records
 - Define semantic relationships among resources to record how objects relate to some external taxonomy or set of standards
 - Model a network overlay where resources are linked together based on contextual information (for example citation links or collaborative annotations)
 - Encode natural hierarchies of objects
 - Make cross-collection linkages among objects (for example show that a particular document in one collection can also be considered part another collection)
- For accessibility and user search and discovery, we will depend on Solr indexing and Blacklight for search. This Ruby on Rails gem offers the following features:
 - faceted browsing (e.g., filtering)
 - relevance based searching (local control of algorithms)
 - bookmarkable items
 - permanent URLs for every item
 - user tagging of items

3. **DIP:** By default, Fedora will store objects internally in their portable form.

- All updates to the Fedora Repository and/or migrations—undertaken and performed by IT staff—will preserve the datastreams (including PIDs) and data objects.

- With proper configuration, Hydra is capable of detecting failures in instances and can redeploy any instance at a new location without human intervention. The layered design of configuration management tools enables separation of recovery processes and the actual service, making the Hydra applicable to a wide range of scenarios. (source: 10.1109/CloudCom.2013.158)

METADATA

There are various options for metadata inclusion. At this time, we are proposing to include two schemas: **Dublin Core**, which is necessary for the Fedora Repository, and **DDI-Lifecycle**, a metadata for the social sciences. Down the road—at the same five-year intervals to check format integrity and preservation sustainability—we might consider adding EAD (Encoded Archival Description) or another standard that incorporates attributes for social media data.

Dublin Core

The OAI Protocol for Metadata Harvesting (OAI-PMH) is a standard for sharing metadata across repositories. Every Fedora digital object has a primary Dublin Core record that conforms to the schema (www.openarchives.org/OAI/2.0/oai_dc.xsd). This metadata is accessible using Fedora's OAI-PMH Provider Interface. The primary elements of this OAI standard are:

- `<element ref="dc:title"/>`
- `<element ref="dc:creator"/>`
- `<element ref="dc:subject"/>`
- `<element ref="dc:description"/>`
- `<element ref="dc:publisher"/>`
- `<element ref="dc:contributor"/>`
- `<element ref="dc:date"/>`
- `<element ref="dc:type"/>`
- `<element ref="dc:format"/>`
- `<element ref="dc:identifier"/>`
- `<element ref="dc:source"/>`
- `<element ref="dc:language"/>`
- `<element ref="dc:relation"/>`
- `<element ref="dc:coverage"/>`
- `<element ref="dc:rights"/>`

(See the Appendix for an example of DC metadata for one of the files under consideration in this set.)

Data Documentation Initiative (DDI)

DDI-Lifecycle attributes and qualities:

- Metadata reuse across the data life cycle
- Metadata-driven survey design
- Complex data, e.g., longitudinal data
- Detailed geographic information
- Multiple languages
- Compliance with other metadata standards like ISO 11179
- Process management and automation

DDI-Codebook definition:

- Codebook is a more light-weight version of the standard, intended primarily to document simple survey data. Originally DTD-based, DDI-C is now available as an XML Schema. The current version of DDI-C is 2.5.

DDI authoring options

Several XML authoring tools are available to facilitate the creation of DDI metadata. With a generic XML editor, the user imports the DDI rules (i.e., the DDI XML Schema) into the software and is then able to enter text for specific DDI elements and attributes. The resulting document is a valid DDI instance or file. There are also DDI-specific tools, such as Nesstar Publisher and Colectica which produce DDI-compliant XML markup automatically.

REQUIRED RESOURCES

- Some technical support will be required to check data integrity and enquire our depositors about checksum information. Time devoted to this support should not exceed \$400-\$500.
- Ingest and management of the dataset will leverage existing Fedora Repository resources of the Digital Library Program at the University of Potato. Technical staff is already implementing and testing a Hydra/Fedora repository stack, which should be operational in early Fall of 2014.
- Technologies and long-term sustainability will be part of the repositories overall budget and the material under consideration should have minimal impact on the budget itself.
- We will rely on automated procedures and software as much as possible. But at a minimum of five-year intervals technical staff will manually check a portion of the repository (generally 10%) for format integrity and preservation sustainability. At this time metadata or additional

datastreams may be added to the objects, but this will be a policy decision with much of the actual XML coding automated and scripted.

- The most significant cost will be in producing metadata, which should take two staff members two to three days to produce. This will cost approximately \$1,000. We hope to receive further information from the depositor before proceeding with ingest and encoding.
- As the data curator responsible for datasets of this type, it will be my responsibility to stay informed about research and repository developments with regard to social media data. As a research library, we may also look toward contributing to standards and policies for social media data collection and preservation.

FURTHER DEVELOPMENT AND ACCESS

For curating a dataset of this kind, I am advising that we consider a few other tools and procedures to include with user access or to incorporate into our Hydra/Fedora stack. These tools would augment search and discovery and open avenues for further development of a social media data collection.

Open Annotation plugin for Fedora

The Open Annotation plugin for Fedora is a content-agnostic web service that allows developers to create, query, retrieve, and serialize annotations using the Fedora Commons repository software to store annotations and their content. The plugin is designed to implement the **Open Annotation Core Data Model** (source: openannotation.org):

- The Open Annotation Core Data Model specifies an interoperable framework for creating associations between related resources, annotations, using a methodology which conforms to the Architecture of the World Wide Web. Open Annotations can easily be shared between platforms, with sufficient richness of expression to satisfy complex requirements while remaining simple enough to also allow for the most common use cases, such as attaching a piece of text to a single web resource.
- An Annotation is considered to be a set of connected resources, including a body and target, and conveys that the body is somehow about the target. The full model supports additional functionality, enabling semantic tagging, embedding content, selecting segments of resources, choosing the appropriate representation of a resource and providing styling hints for consuming clients.

csvkit 0.7.3 (beta)

csvkit is a suite of utilities for converting to and working with CSV, the “king of tabular file

formats.” It is inspired by pdftk, gdal and the original csvcut utility by Joe Germuska and Aaron Bycoffe. csvkit, the developer claims, is to tabular data what the standard Unix text processing suite (grep, sed, cut, sort) is to text. As such, csvkit adheres to the Unix philosophy.

- Small is beautiful.
- Make each program do one thing well.
- Build a prototype as soon as possible.
- Choose portability over efficiency.
- Use software leverage to your advantage.
- Use shell scripts to increase leverage and portability.
- Avoid captive user interfaces.
- Make every program a filter.
- As there is no formally defined CSV format, csvkit encourages well-known formatting standards.

Fedora Rebuilder Utility (for disaster recovery and data migration). With this utility the entire repository can be rebuilt from the digital object and content files. This capability will be facilitated by:

- Separate daily backups of metadata and datastreams.
- Three copies of the repository will be backed up in three different locations, two of which will be tape back-ups.

JasperReports Library

The JasperReports Library is, according to its developers, the world's most popular open source reporting engine. It is entirely written in Java and it is able to use data coming from any kind of data source and produce documents that can be viewed, printed or exported in a variety of document formats, including HTML, PDF, Excel, OpenOffice and Word.

APPENDIX: DUBLIN CORE METADATA EXAMPLE

```
<?xml version="1.0" encoding="UTF-8"?>
<schema targetNamespace="http://www.openarchives.org/OAI/2.0/oai_dc/"
  xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="qualified" attributeFormDefault="unqualified"
  <annotation>
    <documentation>
      XML Schema 2002-03-18 by Pete Johnston.
      Adjusted for usage in the OAI-PMH.
      Schema imports the Dublin Core elements from the DCMI schema for unqualified Dublin Core.
      2002-12-19 updated to use simpledc20021212.xsd (instead of simpledc20020312.xsd)
    </documentation>
  </annotation>
  <import namespace="http://purl.org/dc/elements/1.1/"
    schemaLocation="http://dublincore.org/schemas/xmls/simpledc20021212.xsd"/>
  <element name="dc" type="oai_dc:oai_dcType"/>
  <complexType name="oai_dcType">
    <choice minOccurs="0" maxOccurs="unbounded">
      <dc:title>shows.csv</dc:title>
      <dc:creator>Giglietto, Fabio</dc:creator>
      <dc:creator>Selva, Donatella</dc:creator>
      <dc:subject>Twitter</dc:subject>
      <dc:subject>politics</dc:subject>
      <dc:subject>talk show</dc:subject>
      <dc:subject>big data</dc:subject>
      <dc:subject>Italy</dc:subject>
      <dc:description>Aggregation by television show and tweets of responses to and discussion of political
        elections</dc:description>
      <dc:description>Data aggregated with a peaks detection algorithm</dc:description>
      <dc:publisher></dc:publisher>
      <dc:contributor></dc:contributor>
      <dc:date>2012-2013</dc:date>
      <dc:type>dataset</dc:type>
      <dc:format>comma separated value file</dc:format>
      <dc:identifier></dc:identifier>
      <dc:source>Twitter firehose</dc:source>
      <dc:source>DiscoverText GNIP importer</dc:source>
      <dc:language>Eng</dc:language>
      <dc:relation>SECOND SCREEN AND PARTICIPATION: A CONTENT ANALYSIS ON A FULL SEASON
        DATASET OF TWEETS</dc:relation>
      <dc:relation>Journal of Communication 64.2 (April 2014): 260–277</dc:relation>
      <dc:coverage>From 30th of August 2012 to 30th June 2013 2,489,669 tweets collected</dc:coverage>
      <dc:rights></dc:rights>
    </choice>
  </complexType>
</schema>
```

Notes for Article References

¹ Burgess, J., & A. Bruns (2012, Oct.). Twitter Archives and the Challenges of "Big Social Data" for Media and

Communication Research. *M/C Journal*, 15(5). Retrieved 19 Jun. 2014, from <http://journal.media-culture.org.au/index.php/mcjournal/article/view/561>

² body, dana. "Making Sense of Privacy and Publicity." SXSW Keynote. March 13, 2010. Retrieved 19 Jun. 2014, from <http://www.danah.org/papers/talks/2010/SXSW2010.html>

³ Giglietto, Fabio and Selva, Donatella, Second Screen and Participation: A Content Analysis of a Full Season Dataset of Tweets (October 25, 2013). Available at SSRN: <http://ssrn.com/abstract=2345240> or <http://dx.doi.org/10.2139/ssrn.2345240>. Retrieved 15 Jun. 2014, from <http://ssrn.com/abstract=2345240>

⁴ Giglietto, F. and Selva, D. (2014), Second Screen and Participation: A Content Analysis on a Full Season Dataset of Tweets. *Journal of Communication*, 64: 260–277. doi: 10.1111/jcom.12085