News Production Workflows in Data-driven, Algorithmic Journalism
A Systematic Literature Review

Julian Ausserhofer\textsuperscript{1,2}, Robert Gutounig\textsuperscript{1} & Michael Oppermann\textsuperscript{2}
@julauss, @sextus_empirico, @oppermann_m

1: FH Joanneum University of Applied Sciences Graz
2: University of Vienna

Dubrovnik Media Days, 31.10.15
#DuMD2015 & #DUMediaDays
Research Interest
"experimental use of algorithms, data and social science methods" in journalism (Gynnild, 2014, p. 715)
Computational methods in journalism

(Meyer, 1973)

1952

(Cox, 2000)

http://www.computerhistory.org/revolution/early-computer-companies/5/102/445
Computer-Assisted Reporting: A National Use Update

Bruce Garrison
School of Communication
University of Miami
bgarrison@miami.edu

Society of Professional Journalists national convention
September 19, 1996, Washington, D.C.

Basic Hardware
- Average hard drive size is 1.66 GB
- 9% use Macintosh
- 33% use Pentium PCs
- 51% use DOS/Windows
- 15% use Windows 95 or Windows NT

Online Services, 1996
- 81% use online services
- 37% use online services daily
- 63% use them weekly or more often
- 43% assign reporters or anyone in newsroom to conduct searches

(Garrison, 1996)
The datafication of journalism

"datafication" of society
(Mayer-Schönberger & Cukier, 2013)

"increased focus on measurement, outcomes assessment"
(Anderson, 2015, p. 363)

"journalism's quantitative turn"
(Coddington, 2015)

... has also affected the "infrastructures of journalism"
(Russ-Mohl, 2006, p. 199)
Research literature on data journalism

"internalist tendencies at [... the] early stage of academic research" (Anderson, 2013, p. 1007)

↓

"an explosion in data journalism-oriented scholarship" (Fink & Anderson, 2015, p. 476)*

"rapidly growing body" of scientific studies (Lewis, 2015, p. 322)*

*cited via (Loosen, Reimer & Schmidt, 2015, p. 2)
Research questions

In what way have journalistic workflows and routines changed due to the broad introduction of data in the newsroom?

How has this impacted journalistic norms and ethics, and the skill requirements for journalists?
Method
Systematic literature review
Why a systematic literature review?

"to develop insights, critical reflections, future research paths and research questions"
(Massaro, Dumay & Guthrie, forthcoming)

It adopts "a replicable, scientific and transparent process [...] that aims to minimize bias [...]"
(Tranfield, Denyer & Smart, 2003)
Undertaking a systematic literature review

Writing a literature review protocol

- Defining the questions that the literature review should answer
- Determining the type of studies and carrying out a comprehensive literature search
- Coding data
- Developing insights and critique through analyzing the dataset
- Developing future research paths and questions

Adapted from Massaro et al. (forthcoming)
Own visualization

@julauss, @sextus_empirico & @oppermann_m
In **what way have journalistic workflows and routines changed** due to the broad introduction of data in the newsroom?

How has this impacted **journalistic norms and ethics**, and the **skill requirements** for journalists?
Empirical research on DDJ
Social science focus, but open to other disciplines
Published after 1995

<table>
<thead>
<tr>
<th>Included</th>
<th>Not included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal articles</td>
<td>Bachelor's and Master's theses</td>
</tr>
<tr>
<td>Book sections</td>
<td>Press reports</td>
</tr>
<tr>
<td>Conference papers</td>
<td>Blog posts</td>
</tr>
<tr>
<td>Reports (from industry and research projects)</td>
<td></td>
</tr>
<tr>
<td>PhD theses</td>
<td></td>
</tr>
</tbody>
</table>
Preliminary search with “data-driven journalism”

- Extracting related terms from the keyword section of research papers
Developing insights and critique through analyzing the dataset

Coding data

Developing future research paths and questions

Defining the questions that the literature review should answer

**Search terms**

- algorithmic journalism
- computational journalism
- computer-assisted reporting
- data journalism
- data-driven journalism
- data-driven reporting
- database journalism
- datajournalism
- datenjournalismus
- quantitative journalism

**No search terms**

- accountability journalism
- crowdsourced journalism
- dataviz
- datavis
- ddj
- drone journalism
- investigative journalism
- online journalism
- open journalism
### Scientific Databases

<table>
<thead>
<tr>
<th>Database</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACM Digital</td>
<td>Sowiport</td>
</tr>
<tr>
<td>EBSCO</td>
<td>Springer</td>
</tr>
<tr>
<td>IEEE</td>
<td>SpringerLink</td>
</tr>
<tr>
<td>JSTOR</td>
<td>Taylor &amp; Francis Online</td>
</tr>
<tr>
<td>ProQuest</td>
<td>Web of Science</td>
</tr>
<tr>
<td>Science Direct</td>
<td>Wiley</td>
</tr>
<tr>
<td>Scopus</td>
<td>Google Scholar</td>
</tr>
<tr>
<td>Sociological Abstracts</td>
<td></td>
</tr>
</tbody>
</table>
772 search results

Assessment of title, abstract & keywords

- by two independently working researchers
  (Thomas et al., 2004)

33 research publications
Determining the type of studies and carrying out a comprehensive literature search

Defining the questions that the literature review should answer

Developing insights and critique through analyzing the dataset

Developing future research paths and questions

Coding data

qualitative coding

computational analysis
Results
Literature Corpus
Research literature on workflows in DDJ
Development of literature over time

n=33
Research literature on workflows in DDJ

Publication type

- Book Section
- Report
- Conference Paper

- Parasie & Dagiral, 2013
- Lewis & Usher, 2013
- Karlsen & Stavelin, 2014
- Diakopolus, 2015

n=33

node size = number of citations in Google Scholar
Research literature on workflows in DDJ
Publication year of referenced papers
1787-2015

Constitution of the United States

n=1151
Research literature on workflows in DDJ
Publication year of referenced papers
1999-2015

2013: 176 cited papers

n=1004
Research literature on workflows in DDJ
Most-cited references

<table>
<thead>
<tr>
<th>Title</th>
<th>Referenced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision Journalism: A Reporter's Introduction to Social Science Methods</td>
<td>9</td>
</tr>
<tr>
<td>Accountability through Algorithm: Developing the Field of Computational Journalism</td>
<td>7</td>
</tr>
<tr>
<td>Open Source and Journalism: Toward New Frameworks for Imagining News Innovation</td>
<td>7</td>
</tr>
<tr>
<td>The Data Journalism Handbook</td>
<td>7</td>
</tr>
<tr>
<td>The Journalist As Programmer: A Case Study of The New York Times Interactive News Technology Department</td>
<td>7</td>
</tr>
<tr>
<td>The Promise of Computational Journalism</td>
<td>7</td>
</tr>
<tr>
<td>Computational journalism</td>
<td>6</td>
</tr>
</tbody>
</table>
Research literature on workflows in DDJ

Corpus and most-cited references

edges = 165

corpus publications, n=33

references, n = 70

size = nr. of citations from the corpus
## Method(s) of data collection

<table>
<thead>
<tr>
<th>Method</th>
<th>Nr. of studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey</td>
<td>3</td>
</tr>
<tr>
<td>In-depth interviews</td>
<td>22</td>
</tr>
<tr>
<td>(Participant) observation</td>
<td>4</td>
</tr>
</tbody>
</table>

@julauss, @sextus_empirico & @oppermann_m
Geographical scope

<table>
<thead>
<tr>
<th>Country</th>
<th>Nr. of studies</th>
<th>Multinational (&gt;4)</th>
<th>Nr. of studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>9</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>9</td>
<td>Norway</td>
<td>1</td>
</tr>
<tr>
<td>Germany</td>
<td>4</td>
<td>Netherlands</td>
<td>1</td>
</tr>
<tr>
<td>Sweden</td>
<td>2</td>
<td>Argentina</td>
<td>1</td>
</tr>
<tr>
<td>Switzerland</td>
<td>2</td>
<td>Austria</td>
<td>1</td>
</tr>
<tr>
<td>Finland</td>
<td>2</td>
<td>Belgium</td>
<td>1</td>
</tr>
</tbody>
</table>
Qualitative analysis -
News Production Workflow
News production workflow
Collecting

Collecting digitalized information from

- publicly available data sets
- requested data
- own research or data collected through crowdsourcing
- automated research (scraping)
- leaked data

(17; Loosen, 2015)
News production workflow

Collecting

*Type of data source*

- government or public bureau
- private corporation
- other organizations like NGOs, research institutes

(28, Loosen, 2015)
News production workflow

Collecting

Access to information
in Norway and USA generally considered good (due to freedom of information legislation)

Barriers to easy data access

- fees
- lacking competencies for exporting data at public offices
- the form data is often delivered (e.g. as PDF)
- lack of legal resources to get access to data that is denied to journalists in the first place
News production workflow

Collecting

Selecting data
Because data journalists don’t have much time in the newsroom, they prefer easy-accessible data, e.g. datasets from public bodies.

They are very cautious when it comes to work with data from companies or private sources.
News production workflow

Analyzing

Methods

● Journalists use statistical programs (e.g. MySQL, Excel, Access, SPSS, Google Docs, Maps, Google Refine, Google Fusion Tables etc.)
● Statistical methods and packages including cluster, network & regression analysis

--> Journalists are quite cautious with it, since they consider this outside of their field of expertise.
News production workflow
Building and visualizing

Results of DDJ research depend in increasing manner on (graphical) display of (automated) data analysis software

Tools and programming languages

- Developers use Python, Ruby on Rails, JavaScript, HTML
- Used third-party software includes MySQL, Access, Excel, Caspio, Tableau, Arcview, TimelineSetter etc.

@julauss, @sextus_empirico & @oppermann_m

#DuMD2015 & #DUMediaDays
News production workflow

Building and visualizing

DDJ is connected to a special way of representation which consists of:

- displaying key messages as graphics and/or
- building interactive web applications

*Storytelling is done through*

(1) interactive web applications with continuous update of data
(2) interactive application based on a stable data set
(3) continuously updated info graphic or
(4) static infographics
News production workflow

Collaboration

- Working alone at small and medium-sized newspapers, and part of teams at larger news organizations
- New connections between journalists and programmers
- Members of teams usually have complementary backgrounds

(6, 9, 14, 18)

#DuMD2015 & #DUMediaDays
News production workflow

Collaboration

- Rise of "annotative journalism"
- Regarding the collaboration between newsroom and ICT-department in media corporations data journalists perceive the relation as troublesome.

Rather data journalists try to bypass ICT department when installing extra software or setting up databases or servers.
News production workflow

Publishing the product

- The goal of DDJ activities is by definition the [very accurate] journalistic publication
- Publication of data sources and raw data following open data principles is mentioned too as an integral part of DDJ.

(6, 17, Coddington, 2014)
Backgrounds of data journalists
Backgrounds of data journalists

- Data journalists are appx. 5 years younger than the average age of German journalists
- All of them have started a university program, 80% have also finished it
- There is no generalizable “data journalism” career path. Background is mostly either social science or informatics.
- Backgrounds include programming, design, typography, info graphics, usability, databases, Web and journalism
- USA: dominance of National Institute for Computer-Assisted Reporting (NICAR)
Enabling and Hindering Factors
Enabling and hindering factors

Critical factors

- Skills of the involved journalists
- Support networks (internal and external) important for the development of skills
- Availability of data
- Tools
- Legal resources
- Available resources
Conclusion
• New kind of people arriving in the newsroom
• New skills needed in the newsroom
• Increase of collaborative newswork
• Shift to data-driven approaches (and epistemologies debates connected to that)
Research gaps

Scope:

- Practices outside of Europe and North America
- The construction of "facts"
- Gender aspects

Methods of data collection:

- Survey
- (Participant) observation
- Content analysis
- Digital methods
Explore the (preliminary) corpus online – it's a prototype:
http://tinyurl.com/valid-survey
http://www.validproject.at

Julian Ausserhofer\textsuperscript{1,2}, @julauss
julian.ausserhofer@fh-joanneum.at

Robert Gutounig\textsuperscript{1} & @sextus_empirico
robert.gutounig@fh-joanneum.at

Michael Oppermann\textsuperscript{2} @oppermann_m
michael.oppermann@univie.ac.at

1: FH Joanneum University of Applied Sciences Graz
2: University of Vienna

Dubrovnik Media Days, 31.10.15
#DuMD2015 & #DUMediaDays
Tools & Literature
Tools & software libraries


Primary Sources


Primary Sources


Corpus 1/3


#DuMD2015 & #DUMediaDays @julauss, @sextus_empirico & @oppermann_m


#DuMD2015 & #DUMediaDays @julauss, @sextus_empirico & @oppermann_m
Explore the (preliminary) corpus online – it's a prototype:
http://tinyurl.com/valid-survey
http://validproject.at

Julian Ausserhofer\textsuperscript{1,2}, @julauss
julian.ausserhofer@fh-joanneum.at

Robert Gutounig\textsuperscript{1} & @sextus_empirico
robert.gutounig@fh-joanneum.at

Michael Oppermann\textsuperscript{2} @oppermann_m
michael.oppermann@univie.ac.at

1: FH Joanneum University of Applied Sciences Graz
2: University of Vienna

Dubrovnik Media Days, 31.10.15
#DuMD2015 & #DUMediaDays