INJECT | FUTURE JOURNALISM, NOW.

**Project Number 732278**

## Deliverable N°: D5.1

# Data Management Plan

## Date

**Version N° 2**

Main Author(s): Professor Neil Maiden, Dr Amanda Brown

Institution(s): City, University of London

Project funded by the European Union from the EU
H2020 Programme under grant agreement number 732278

Horizon 2020
European Union Funding
for Research & Innovation

| | |
|---|---|
| **Project ref. no.** | 732278 |
| **Project title** | INJECT: Innovative Journalism: Enhanced Creativity Tools |

| | |
|---|---|
| **Nature of Deliverable** | **R** = Report |
| **Contractual date of delivery** | Project Month Six – June 2017 |
| **Actual date of delivery** | 30 June 2017 |
| **Deliverable number** | D5.1 |
| **Deliverable title** | Data Management Plan |
| **Dissemination Level** | Public |
| **Status & version** | Final 2.0 |
| **Number of pages** | 17 |
| **WP relevant to deliverable** | WP5 |
| **Lead Participant** | City, University of London |
| **Author(s)** | Professor Neil Maiden, Dr Amanda Brown |
| **Project coordinator** | Neil Maiden, City University London, UK |
| **EC Project Officer** | Albert Gauthier |
| **Keywords** | INJECT, Data Management, FAIR, Journalism |

# Table of Contents



# Table of figures

# Executive Summary

INJECT is a new Innovation Action that supports technology transfer to the creative industries; under the call for “action primarily consisting of activities directly aiming at producing plans and arrangements or designs for new, altered or improved products, processes or services” (H2020 Innovation Action). To achieve its aim INJECT will test and establish an INJECT spin-off business in the journalism market through its ecosystem developments. While user testing and testing of the tool in operational environments will aid in the development and technical improvements of the INJECT technology.

The INJECT tool is new to journalism and to European markets; the data management plan covers this testing and validation of both technical and economic performance in real life operating conditions provided by the journalism market domain. This project therefore has limited scientific research activities.

This document aims to present the data management plan for INJECT project, the considerations, actions and activities planned with an aim to deliver on the objectives of the project. The deliverable introduces the data management plan as a living document, its purpose and intended use. The document discusses the INJECT data types and applies the FAIR data management process to ensure that, wherever possible, the research data is findable, accessible, interoperable and reusable (FAIR), and to ensure it is soundly managed.

# 1 Purpose of the Data Management Plan

This deliverable of the INJECT project is prepared under WP5 and the Task 5.1 *INJECT Data Management Plan (1st version)*. In this task we initiate discussion of the data management life cycle, processes and/or generated by the INJECT project and to make the data findable, accessible, interoperable and reusable (FAIR). This data management plan is living document, a dynamic document that will be edited and updated during the project, with a second version to be delivered in month 18.

# 2 INJECT Data Types

The Data Management Plan asks the following questions and we address those throughout the document, noting where actions are underway and further considerations that will be made as the project develops. As previously noted, the INJECT project is an H2020 Innovation Action, and hence is not intended to generate scientific data per se, therefore the data management plan considers the activities undertaken within the project.

## 2.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

The three stated INJECT project objectives are:

Obj1: Extend and aggregate the new digital services and tools to increase the productivity and creativity of journalists in different news environments

Obj2: Integrate and evaluate the new digital services and environments in CMS environments

Obj3: Diffuse the new digital services and support offerings in news and journalism markets

Data collection and generation related to each is to enable the co-creation then effective evaluation of the INJECT tool, and scientific reporting of research and innovation that will deliver each of these objectives.

### 2.1.1 What types and formats of data will the project generate/collect?

The project will generate and collect the following types and formats of data:

- Co-created user requirements on the INJECT tool and services: format is structured text requirements;
- Parsed and semantic-tagged news stories from online digital news sources (including partner news archives) as part of INJECT toolset: format is PostgreSQL database, the processed/parsed results are stored into an external Elastic Search Cluster for later searching;

- Semantic-tagged news stories used to inform design of INJECT creative search strategies: format is structured documents of news stories, associated word counts and other observed patterns, by story type;
- Usability evaluation reports of INJECT tool by journalists: format is structured written reports;
- Semi-structured interview data about INJECT tool use by journalists: format is documented, content-tagged notes from semi-structured interviews;
- Focus group reports about INJECT tool use by journalists: format is structured reports of focus group findings;
- INJECT tool activity log data, recording meaningful activities of tool users over selected time periods: format is structured spreadsheet;
- Corpus of news stories generated by journalists using the INJECT tool: format is structured database of news stories and related data attributes;
- Quantitative creativity assessments of news stories generated by journalists with and without use of the INJECT tool: format will be structured spreadsheets;
- Economic and contract data about each launched INJECT ecosystem: format is structured spreadsheet.

### 2.1.2 Will you re-use any existing data and how?

The following data is reused from existing news sources:

- Parsed and semantic-tagged news stories from online digital news sources (including partner news archives) as part of INJECT toolset: format is the raw news article data is stored in a PostgreSQL database, the processed/parsed results are stored into an external Elastic Search Cluster for later searching;
- Semantic-tagged news stories used to inform design of INJECT creative search strategies: format is structured documents of news stories, associated word counts and other observed patterns, by story type;
- Corpus of news stories generated by journalists using the INJECT tool: format is structured database of news stories and related data attributes.

### 2.1.3 What is the origin of the data?

The reused data originates from selected news sources:

*Figure 1: News Sources*

| Source | Country |
|---|---|
| BBC | UK |
| Quartz | UK |
| The Guardian | UK |
| Telegraph | UK |
| FT | UK |

| | |
|---|---|
| The Times | UK |
| Sky News | UK |
| The Independent | UK |
| The Huffington Post | UK |
| The Huffington Post | US |
| Reuters News | UK |
| The Economist | UK |
| The New York times | US |
| Daily Mail | UK |
| The Wall Street Journal | US |
| The Washington Post | US |
| The Metro | UK |
| Herald Scotland | UK |
| Bloomberg | US |
| The Scotsman | UK |
| The Irish Times | Ireland |
| Irish Independent | Ireland |
| New Statesman | UK |
| Newsweek | US |
| The Daily Beast | US |
| Times Education Supplement | UK |
| BBC Mundo | UK |
| El Mundo | Spain |
| El Pais | Spain |
| Cinco Dias | Spain |
| CNN | US |
| CNN Money | US |
| London Evening Standard | UK |
| Birmingham Post | UK |
| Birmingham Mail | UK |
| Farming Life | UK |
| Belfast Telegraph | UK |
| Yorkshire Post | UK |
| Yorkshire Evening Post | UK |
| Manchester Evening News | UK |
| South Wales Evening Post | UK |
| Irish Examiner | Ireland |
| Herald Scotland | Scotland |
| The Mirror | UK |
| The Irish Sun | Ireland |
| Irish Daily Star | Ireland |
| The Sun | UK |

| | |
|---|---|
| Daily Star | UK |
| Daily Record | UK |
| Daily Express | UK |
| Los Angeles Times | US |
| Chicago Tribune | US |
| The Onion | US |
| Forbes | US |
| Fox News | US |
| Herald Tribune [International NY Times] | US |
| ABC News | US |
| Buzzfeed | US |
| Newsmax Media | US |
| U.S. News and World Report | US |
| The Globe and Mail | Canada |
| Toronto Star | Canada |
| New Zealand Herald | NZ |
| Dominion Post | NZ |
| The Sydney Morning Herald | Australia |
| The Brisbane Times | Australia |
| Herald Sun | Australia |
| The Daily Telegraph (Australia) | Australia |
| The Courier-Mail | Australia |
| Bangkok Post | Thailand |
| Jakarta Globe | Indonesia |
| South China Morning Post | Hong Kong |
| Der Spiegel International | Germany |
| Ekathimerini | Greece |
| Dutch News | Netherlands |
| Krakow Post | Poland |
| Portugal Resident | Portugal |
| The Local Newspaper | Sweden |
| Connexion Newspaper | France |
| Le Monde | France |
| Le Monde Diplomatique | France |
| EuroFora | EU |
| Friedl News | Austria |
| New Europe | Belgium |
| Copenhagen Post | Denmark |
| News of Iceland | Iceland |
| Finnbay Newspaper | Finland |
| North Cyprus News | Cyprus |
| Prague Daily Monitor | Czech Republic |

| | |
|---|---|
| Daily News Egypt | Egypt |
| The Punch | Nigeria |
| Business Day Live | South Africa |
| Independent Newspaper | South Africa |
| Mail and Guardian | South Africa |
| Bhutan Observer | Bhutan |
| Financial Express | India |
| Business Standard | India |
| Economic Times | India |
| The Indian Express | India |
| Live Mint [INDIA] | India |
| Stavanger Aftenblad | Norway |
| Bergens Tidende | Norway |
| Dagbladet | Norway |
| Verdens Gang (VG) | Norway |
| Dagens Næringsliv | Norway |
| NRK | Norway |
| Aftenposten | Norway |
| Le Figaro | France |
| BFMTV | France |
| Le Parisien | France |
| Le Express | France |
| L'OBS | France |
| Le Point | France |
| Les Echos | France |
| CBS | Netherlands |
| SCP | Netherlands |
| NU | Netherlands |
| Al Jazeera | Qatar |
| FD | Netherlands |
| Adformatie | Netherlands |
| Eerste Kamer | Netherlands |
| Europees Parlement Nieuws | Netherlands |
| Daily Nation | Kenya |
| Vanguard | Nigeria |
| The Namibian | Namibia |
| News24 | South Africa |

As the first ecosystem for INJECT is established in Norway there will be more sources that may be added, such as internal archives, statistical bureau information, and public data (maps, weather, traffic). It is further noted that this list will expand with further ecosystem developments as more newspapers and others from the journalistic domain became customers in the future.

### 2.1.4 Data generated during the project arises from:

- A user-centred co-design process with journalists and news organisations;
- Knowledge acquisition and validation exercises with experienced journalists for each of the 6 INJECT creative search strategies;
- Data- and information-led design of each of the 6 INJECT creative search strategies;
- Formative and summative evaluations of INJECT tool use by journalists and news organisations.
- Original content created by journalists and news organisations who choose to contribute to public Explaain card content.

### 2.1.5 What is the expected size of the data?

The expected sizes of the data varies by types:

- Documents and reporting describing the user requirements, user activity logs and qualitative results from formative and summative evaluations of the INJECT tool, including the corpus of generated news stories, will be small – deliverable reports with short data appendices;
- Parsed and semantic-tagged news stories from online digital news sources (including partner news archives) as part of INJECT toolset will be large. The current data set at m6 of the project is just over one million articles.

### 2.1.6 To whom might it be useful ('data utility')?

The INJECT project data might be useful to:

- News organisations and IT providers who will target the news industry, to inform their development of more creative and productive news stories, to support the competitiveness of the sector;
- News organisations and IT providers who wish to develop new forms of business model through which to deliver digital technologies to the news and journalism sectors;
- Journalism practitioners who will extrapolate from project results in order to improve journalism practices across Europe.
- Academics and University departments and Institutes that could use the INJECT data for research and teaching purposes.

# 3 FAIR data

## 3.1 Making data findable, including provisions for metadata

As stated previously INJECT is an Innovation Action that supports technology transfer to the creative industries; it will test and establish an INJECT spin-off business in the journalism market through its ecosystem developments. The INJECT tool is new to journalism and to European markets and the intention is that it becomes a sought after commercially viable product. This viability will require the product to be sold and to earn revenue, from both its subscribed use and innovations made through

paid for adaptations. It will be necessary that some types of information are sold specifically to customers and therefore cannot be in the public domain.

The FAIR framework asks:

- Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?
- What naming conventions do you follow?
- Will search keywords be provided that optimize possibilities for re-use?
- Do you provide clear version numbers?
- What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

The following table provides INJECT's current answers to these questions.

*Figure 2: Making data findable.*

| **Data type** | **Discoverable?** | **Reuse and metadata conventions** |
|---|---|---|
| Co-created user requirements on the INJECT tool and services | No | The single user requirements document will be extracted from project deliverables, and posted in an acceptable form on the INJECT project website |
| Parsed and semantic-tagged news stories from online digital news sources as part of INJECT toolset | Yes | All news stories will be searchable through the INJECT tool and advanced search algorithms, which have APIs. News stories are tagged with semantic metadata about article nouns and verbs, and person, place, organisation and activity entities. The meta-data types are currently bespoke standards, to allow tool development to take place |
| Semantic-tagged news stories used to inform design of INJECT creative search strategies | No | The news stories will be collated in one or more online documents. Each news article will be meta-tagged with data about the article's length, presence and number of keywords, and other observations |
| Usability evaluation reports of INJECT tool by journalists | No | The usability evaluation report content will not be made available for reuse. Ethical approval does not allow for reuse and sharing |
| Semi-structured interview data about INJECT tool use by journalists | No | The semi-structured interview data will not be made available for reuse, as ethical approval does not allow for its reuse and sharing |

| | | |
|---|---|---|
| Focus group reports about INJECT tool use by journalists | No | The focus group data will not be made available for reuse, as ethical approval does not allow for its reuse and sharing |
| INJECT tool activity log data, recording meaningful activities of tool users over selected time periods | Yes | Anonymous INJECT tool activity log data will be made available for sharing and reuse, in line with ethical consent from journalist users. Clear log data versions will be set up. Data will be structured and delivered in XLS sheets, to allow analyst searching and management of the data |
| Corpus of news stories generated by journalists using the INJECT tool | No | The corpus of news stories will not be made available directly for reuse by the project, although published articles will be available, at their publication source |
| Quantitative creativity assessments of selected news stories generated by journalists with and without use of the INJECT tool | Yes | Anonymous quantitative creativity assessments of selected news stories generated with and without the INJECT tool will be made available for sharing and reuse, in line with ethical consent from the expert assessors. Clear log data versions will be set up. Data will be structured and delivered in XLS sheets, to allow analyst searching and management of the data |
| Economic and contract data about each launched INJECT ecosystem | No | The intention is that INJECT becomes a sought after commercially viable product to be sold and to earn revenue, from both its subscribed use and innovations made through paid for adaptations. It will be necessary that some types of information are sold specifically to customers and therefore cannot be in the public domain. |

## 3.2 Making data openly accessible

The FAIR framework asks:

- Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.
- How will the data be made accessible (e.g. by deposition in a repository)?
- What methods or software tools are needed to access the data?
- Is documentation about the software needed to access the data included?

- Is it possible to include the relevant software (e.g. in open source code)?
- Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories that support open access where possible.
- Have you explored appropriate arrangements with the identified repository?
- If there are restrictions on use, how will access be provided?
- Is there a need for a data access committee?
- Are there well-described conditions for access (i.e. a machine readable license)?
- How will the identity of the person accessing the data be ascertained?

The following table provides INJECT's current answers to these questions for data that will be made available for sharing in the project.

*Figure 3: Openly accessible data.*

| **Data type** | **Open?** | **How will data be accessed** |
|---|---|---|
| Co-created user requirements on the INJECT tool and services | Yes | The single user requirements document will be posted on the project website, with clear signposting and instructions for use |
| Parsed and semantic-tagged news stories from online digital news sources as part of INJECT toolset | No | The parsed and semantic-tagged news stories will not be made publicly available. This data represents core commercial value of the INJECT tool, and will be not shared, except through INJECT tools made available as part of the commercial ecosystems |
| Semantic-tagged news stories used to inform design of INJECT creative search strategies | Yes | The news stories will be published in online documents that will be accessible via the INJECT's restricted project website and associated storage space. The stories will be stored and edited using standard MS Office applications, which users will need to edit them. A validated user log-in to the restricted area of the INJECT project website will be needed to access and download the stories |
| INJECT tool activity log data, recording meaningful activities of tool users over selected time periods | Yes | The INJECT tool activity log data will be published in online documents that will be accessible via the INJECT's restricted project website and associated storage space. The log data will be stored and edited using standard MS Office applications, which users will need to edit them. A validated user log-in to the restricted area of the INJECT project website will be needed to access and download the log data |
| Quantitative creativity assessments of selected news stories generated by journalists with and | Yes | The collected quantitative assessments will be published in online documents that also will be accessible via the INJECT's restricted project website and associated storage space. The assessments will be stored and edited using standard MS |

| | | |
|---|---|---|
| without use of the INJECT tool | | Office applications, which users will need to edit them. A validated user log-in to the restricted area of the INJECT project website will be needed to access and download the quantitative assessments |
| Economic and contract data about each launched INJECT ecosystem | No | The intention is that INJECT becomes a sought after commercially viable product with innovations made through paid for adaptations. It will be necessary that some types of information are sold/contracted to specific customers and therefore cannot be in the public domain. |

## 3.3 Making data interoperable

The FAIR assessment asks:

- Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?
- What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?
- Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?
- In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

In response, the INJECT project will not seek to make its data interoperable with other research data sets, and to enable data exchange and re-use between researchers, institutions, organisations and countries. There are several reasons for this decision:

- There are no established standards for data about digital tool use in journalism, to interoperate with;
- There are established standards for data about creativity support tool use in computer science, to interoperate with, although a standardized survey metric for digital creativity support has been developed by US researchers, which the INJECT project will submit to.

To compensate, the INJECT project will make its data available in the most open tools available, for example the MS Office suite, and to provide sufficient documentation to enable understanding and use by other researchers.

## 3.4 Increase data re-use (through clarifying licences)

The FAIR framework asks:

- How will the data be licensed to permit the widest re-use possible?
- When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.
- Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.
- How long is it intended that the data remains re-usable?
- Are data quality assurance processes described?

Data re-use is a live consideration for INJECT as the tool is technically developed and ecosystems established. City and the Innovation Manager are leading an exploration into the registrations of one or more trademarks for the project. The current recommended action for public documents, such as the website, have been marked with the copyright symbol (©), name and the year of creation: Copyright © The INJECT Consortium, 2017. Data protection aspects of the project will be coordinated across the relevant national data protection authorities. The project is aware, and will work towards, upcoming European data protection rules that will enter into force May 2018 and their impact will be considered: http://ec.europa.eu/justice/data-protection/reform/index_en.htm

In addition, an ongoing investigation into Intellectual Property rights is underway. Advice is has been sought through legal channels at City, University of London. This includes consideration of how the INJECT tool operates in framing and storing of article text and referencing plus the eco-systems' payment and use of the tool. As the project develops this will be a key consideration in work packages.

## 3.5 Allocation of resources

The FAIR framework asks:

- What are the costs for making data FAIR in your project?
- How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).
- Who will be responsible for data management in your project?
- Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

The FAIR framework has a minimum impact on INJECT. INJECT's resources for managing the FAIR framework are built into the project's work plan. For example:

- The development and management of the INJECT data types and sets is incorporated into and budgeted for in the current work plan;
- Overall data management will be undertaken by the project manager role at the project coordinator partner – Dr Amanda Brown.

However, the resources for long-term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long) have yet to be finalised for the first version of the FAIR document.

## 3.6 Data security

The FAIR framework asks:

- What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?
- Is the data safely stored in certified repositories for long-term preservation and curation?

INJECT stores the processed/parsed results into an Amazon Elastic Search Cluster. Amazon Elasticsearch Service routinely applies security patches and keeps the Elasticsearch environment secure and up to date. INJECT controls access to the Elasticsearch APIs using AWS Identity and Access Management (IAM) policies, which ensure that INJECT components access the Amazon Elasticsearch clusters securely. Moreover, the AWS API call history produced by AWS CloudTrail enables security analysis, resource change tracking, and compliance auditing.

## 3.7 Ethical aspects

The FAIR framework asks:

- Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review.
- Is informed consent for data sharing and long-term preservation included in questionnaires dealing with personal data?

The INJECT consortium have not identified any specific ethics issues related to the work plan, outcomes or dissemination. We do note that individual partners will adhere to ethical rules.
At City, University of London the data management and compliance team are undertaking a significant review of all policies and procedures on ethics and data use. We continue to work to the current data protection policy with a commitment to protecting and processing data with adherence to legislation and other policy. “Sensitive data shall only be collected for certain specific purposes, and shall be obtained with consent” will apply to all personal data collected and any participants provided fair processing notices about the use of that data. The project will adhere to the commitment to holding any data in secure conditions, and will make every effort to safeguard against accidental loss or corruption of data.

# 4 Summary and Outlook

The subsequent INJECT deliverable D5.2 will revisit the data management plan, the considerations, actions and activities undertaken alongside the delivery on the objectives of the project. “The FAIR Data Principles provide a set of milestones for data producers” (Wilkinson et al, 2016) and as the project develops and within the next deliverable we will revisit the data management plan data types and consider the milestones to apply the FAIR data management of research data that is findable,

accessible, interoperable and reusable (FAIR).

# 5 References

European Commission (2014) H2020 The EU Framework Programme for Research & Innovation https://www.upf.edu/rdi/_pdf/H2020_inBrief_EN_FinalBAT.pdf

H2020 *Programme Guidelines on FAIR Data Management in Horizon 2020* http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf July 2016

Wilkinson, M. D. et al. (2016) *The FAIR Guiding Principles for scientific data management and stewardship.* Sci. Data 3:160018 doi: 10.1038/sdata.2016.18