OPINION ARTICLE

Related to other papers in this special issue	2 (p10); 4 (p40); 23 (p230); 9 (p87); 8 (p78)
Addressing FAIR principles	F, A, I, R

Social Data: CESSDA Best Practices

Ron Dekker[†]

Main Office of the Consortium of Social Science Data Archives, Parkveien 20, Bergen, Norway

Keywords: Social data; Distributed infrastructure; FAIR Action Plan; Data catalogue

Citation: R. Dekker. Social data: CESSDA best practices. Data Intelligence 2(2020), 220–229. doi: 10.1162/dint_a_00044

ABSTRACT

The European Commission report "Turning FAIR into reality" provides an index of 27 FAIR Action Plan recommendations. This index is used for a self-assessment on CESSDA, the Consortium of European Social Science Data Archives. CESSDA is performing well on "Concepts for FAIR implementation", "Skills for FAIR", and "Investment in FAIR"; there is work in progress on "FAIR culture", and work to start up on "FAIR ecosystem" and especially on "Incentives and metrics for FAIR data and services". Next, an analysis on the FAIR components, reveals that CESSDA has accomplished the "F", is working on the "A" – considering the sensitivity and security requirements of social data, just started on "I", and that there is lack of clarity on what should be in "R". On Findability, the CESSDA Data Catalogue is explained, showing the building blocks that need to be in place before one can produce a catalogue. The article ends with a forward look on CESSDA's deployment on the FAIR principles.

1. INTRODUCTION

1.1 Social Data Research Infrastructures

The Consortium of European Social Science Data Archives (CESSDA) is a European Research Infrastructure. It has been on the Roadmap of the European Strategy Forum on Research Infrastructures (ESFRI) since 2006

[†] Corresponding author: Ron Dekker (E-mail: ron.dekker@cessda.eu, ORCID: 0000-0003-0989-4963).

and a Landmark since 2014 [2]. In 2017 CESSDA became a European Research Infrastructure Consortium (ERIC) a European legal entity assigned by the European Commission. CESSDA has 19 Members and 1 Observer – and each country must assign a national data service provider. The CESSDA Main Office is in Bergen, Norway[①].

CESSDA's mission is:

- to provide a distributed and sustainable research infrastructure enabling the research community to conduct high-quality research in the social sciences and contributing to the production of effective solutions to the major challenges facing society today; and
- to facilitate teaching and learning in the social sciences.

CESSDA's capital value is 117 M€ and operating costs are 39 M€ per year [2] (p. 212). This makes CESSDA one of the major players within the Social & Cultural Innovation Cluster (see Table 1). The table also shows that this cluster is rather small compared to the other clusters; that it can be characterised as a distributive infrastructure (see also [3] Wittenburg, 2019); and that annual operating costs compared to investment costs are relatively high.

Table 1. Overview of European research infrastructures.

ESFRI-CLUSTER	TOTAL	EOSC-CLS	ERICS	LNDMARKS	PROJECTS	% DISTRIB	CAP VAL M€	OPER CST	% OPER/VAL
DATA	1	0	0	1	0	100	500	60	12
ENERGY	6	0	1	2	4	50	3,354	127	4
ENVIRONMENT	11	11	4	7	4	91	2,298	219	10
HEALTH & FOOD	16	14	6	10	6	100	2,410	337	14
PHYSICAL SCIENCES & ENGINEERING	16	8+6	3	12	4	44	10,336	1,070	10
SOCIAL & CULT. INNOVATION	7	6	5	5	2	100	382	81	21
TOTAL	**57**	**45**	**19**	**37**	**20**	**77**	**19,280**	**1,894**	**10**

Next to its distributive structure and high annual running costs, there is another important feature of social data: the amount of sensitive data. Rough estimates indicate that 40% of the data need protective measures, e.g. further anonymization, remote execution, secured access. This has implications on the degree of openness and accessibility of social data, and may limit interoperability of the data (connecting data with contextual data, linking data using semantics).

[①] www.cessda.eu.

Other characteristics of social data are the multilingualism, requiring additional actions to make data comparable to make the data reusable, and concerns on data quality and provenance: as everybody can generate social data, it is important that there is a quality check on the data. As a first step in providing trust and ensuring data quality, CESSDA requires that its national service providers have the CoreTrustSeal (CTS[2]).

1.2 Open Science and FAIR

Open Science is an umbrella term for trends in scientific communication, data sharing, and software development. A good explanation why we need open science is given in the Lamy-report [4](p. 8): "Europe must embrace the transformative power of open science allowing for a faster circulation of increasing amounts of knowledge, and seize the potential of open innovation to trigger faster and fairer growth, building a knowledge economy that is open to the world".

2. FAIR ACTION PLAN

The FAIR principles [5] are connected to open science, stressing that research data should be Findable, Accessible, Interoperable and Reusable [6]. The EC Expert Group on FAIR Data, chaired by Simon Hodson and Sarah Jones, developed an Index to FAIR Action Plan Recommendations [1] (p. 60).

This index has been used for a self-assessment on CESSDA and is a first attempt to map the readiness of the infrastructure to the FAIR Index. There are 27 recommendations, specifying actions for different stakeholders. In the assessment, each recommendation is scored on recommended activities for service providers:

- Green: activities are taking place in a sufficient way
  – green does not imply that activities can stop;
- Orange: activities are planned (in the pipeline) or just started;
- Red: activities are not planned yet.

The scoring is subjective, based on CESSDA activities (strategy, working plans, projects). Ideally such a self-assessment should be checked by an external team, but as this is a first attempt to use the FAIR Index, we are mainly interested whether this approach is usable at all.

[2] www.coretrustseal.org.

We used the FAIR Action Plan Index to map CESSDA's position (Table 2).

Table 2. Index to FAIR Action Plan Recommendations.

Concepts for FAIR implementation	FAIR culture	FAIR ecosystem	Skills for FAIR	Incentives and metrics for FAIR data and services	Investment in FAIR
1. Define FAIR for implementation	4. Develop Interoperability frameworks	7. Support semantic technologies	10. Profession-alise data science & stewardship roles	12. Develop metrics for AIR Digital Objects	14. Provide strategic and coordinated funding
2. Implement a model for FAIR Digital Objects	5. Ensure data management via DMPs	8. Facilitate automated processing	11. Implement curriculum frameworks and training	13. Develop metrics to certify FAIR services	15. Provide sustainable funding
3. Develop components of a FAIR ecosystem	6. Recognise & reward FAIR data & stewardship	9. Certify FAIR services	Above line = priority recommendations		
16. Apply FAIR broadly	18. Cost data management	22. Use information held in DMPs	Below line = supporting recommen-dations	25. Implement and monitor metrics	27. Open EOSC to all providers but ensure services are FAIR
17 Align and harmonise FAIR and Open Data policy	19. Select and prioritise FAIR digital objects	23. Develop components to meet research needs		26. Support data citation and next generation metrics	
	20. Deposit in Trusted Digital Repositories	24. Incentivise research infrastructures to support FAIR data			
	21. Incentivise reuse of FAIR data				

The Consortium is performing well on "Concepts for FAIR implementation", "Skills for FAIR", and "Investment in FAIR", as most of the recommendations are already taking place (green in the table), or are in the pipeline (orange in the table). There is work in progress on "FAIR culture" (most recommendations are orange, that is, in the pipeline), and work to start up on "FAIR ecosystem" and especially on "Incentives

and metrics for FAIR data and services" (most recommendations did not start (green), but are in the pipeline (orange) or not even started (red)).

In detail the scores are:

CESSDA is performing well on[3]

- "Concepts for FAIR implementation"
  CESSDA Service Providers put a lot of effort into long-term stewardship, timeliness of data sharing, assessability and legal issues (1); CESSDA has a persistent identifiers policy that has been implemented, and uses tools to capture metadata from the national service providers (2); CESSDA distinguishes and uses the essential components of the FAIR ecosystem (3).
  CESSDA is working on case studies, getting into contact with research communities (16), CESSDA is part of Social Sciences and Humanities (SSHOC) and ERIC-Forum, aligns with the DDI (metadata system), is one of the co-founders of the Interest Group on Social Sciences Research Data at the Research Data Alliance, and working on secured access tools (17).
- "Skills for FAIR"
  As training and sharing expertise is in CESSDA's mission, there is a lot of attention and effort into training of researchers and service providers' staff (10). There is a mentorship programme for new service providers and there is an internal quality check before training goes online (11).
- "Investment in FAIR"
  Strategic and coordinated funding (14) seems hardly applicable for service providers, although CESSDA has been discussing the fall-back procedure should one of the national service providers suddenly stop its activities. The setting of being an ERIC provides stability for the funding of the Consortium (15), although there remain concerns on the national support for some service providers. Ensuring that services are FAIR (27) relates to CESSDA's quality assessment of its national service providers (including CoreTrustSeal), but also to strive for full European coverage and strengthening new service providers (via seminars, consortium meetings and mentorships).

There is work in progress on

- "FAIR culture"
  CESSDA is active on making data interoperable in cooperation with research communities and new platforms or market places for sharing tools and data are being developed in the European EOSC-cluster project SSHOC (4). Service Providers encourage the use of Data Management Plans (DMPs) and CESSDA has a widely used DMP course online (5). CESSDA started up a task force on data citation (6), which is a prerequisite for acknowledgement on data sharing.
  On cost data management (18) there are ideas, but no activities yet. Selecting and prioritising FAIR Digital Objects are at the core activities of national service providers (19), and the same goes for Deposit in Trusted Digital Repositories (20) – again in cooperation with research communities.

[3] The numbers in brackets refer to the FAIR Index Recommendations (Table 2).

Reaching out to users and encourage and incentivise reuse of FAIR outputs must start (21) – for this we first need to extend connections with the communities.

and work to start up on

- "FAIR ecosystem"
  CESSDA has participated in semantic data projects and this is also being addressed in the 14.5 M€ SSHOC project. Service providers work on new data types (e.g. social media data, registries) to be combined or linked with surveys (7). Automated processing (8) is a goal, but must start yet – first CESSDA needs to optimise its workflows within the FAIR ecosystem. CESSDA is very active on certification (9), including the CoreTrustSeal and developing internal assessment procedures.
  On using information held in DMPs there must first be increased use of DMPs to have more content (22); CESSDA is working on metadata profiles for different parts of the DMPs, e.g., on data sets, variables, and questionnaires. Extension of FAIR components (23) has not started yet: priority is to work on the existing ones. CESSDA is active on supporting FAIR data – also in cooperation within the Social Sciences and Humanities cluster (SSHOC project) and within the ERIC-Forum (24).

and especially on

- "Incentives and metrics for FAIR data and services"
  Metrics on FAIR Digital Objects (12) and monitoring (25) are still to do, preferably with other partners (in SSHOC or ERIC-Forum). Certification of FAIR Services (13) has some work in progress, extending on the CoreTrustSeal. On data citation (metrics) (26) CESSDA has started a task force.

2.1 Conclusion

The FAIR Index [1] has been used to monitor research infrastructures' status on implementation of FAIR policies. After a zero measurement, the index can be used to measure progress: how does the mapping change in a year?

We used only three categories and used colours instead of scores (green: implemented; orange: in the pipeline; red: not yet started). One could be tempted to make the scoring more quantitative, just to bring in more objectivity or to be able to add the scores and construct one single number. However, adding up scores would give false accuracy and scores might become a goal in itself. Using colours gives a clear dashboard and quick status overview.

This index should not be about reaching high score, but about making progress – where are the weak spots and is there room for improvements. After the monitoring the organisation should set up an Action Plan with concrete actions to get more and more green marks and diminishing the red ones. This Indexing should be a helpful tool and not a goal on itself.

2.2 Strategy on FAIR – start with Findability

The FAIR Action Plan Index describes FAIR actions, but does not distinguish between F, A, I, and R. In the CESSDA Strategy the FAIR principles are boundary conditions, but we also opted to have a sequential approach: First focus on Findability, which is without prejudice to the fact that we are also working on the other principles.

In this section, we will present how CESSDA worked towards this Findability, by implementing a Data Catalogue (Figure 1) with metadata on data sets (studies) from all the national service providers.

The good news: the CESSDA Data Catalogue has been online since the end of 2018. The bad news: it took over three years to produce it and we had to build some back-office tools first.

The catalogue holds over 25,000 Studies (19,000 in English, and other languages also available) and besides free text search (in basic and advanced mode), it allows filtering on language, topic, years, country, publisher (service provider), and language of data files. It is fast and harvests the metadata every night. As the catalogue contains only metadata there are no privacy or security issues with the data. A user can click on a Study to go directly to the Service Provider to access the data.

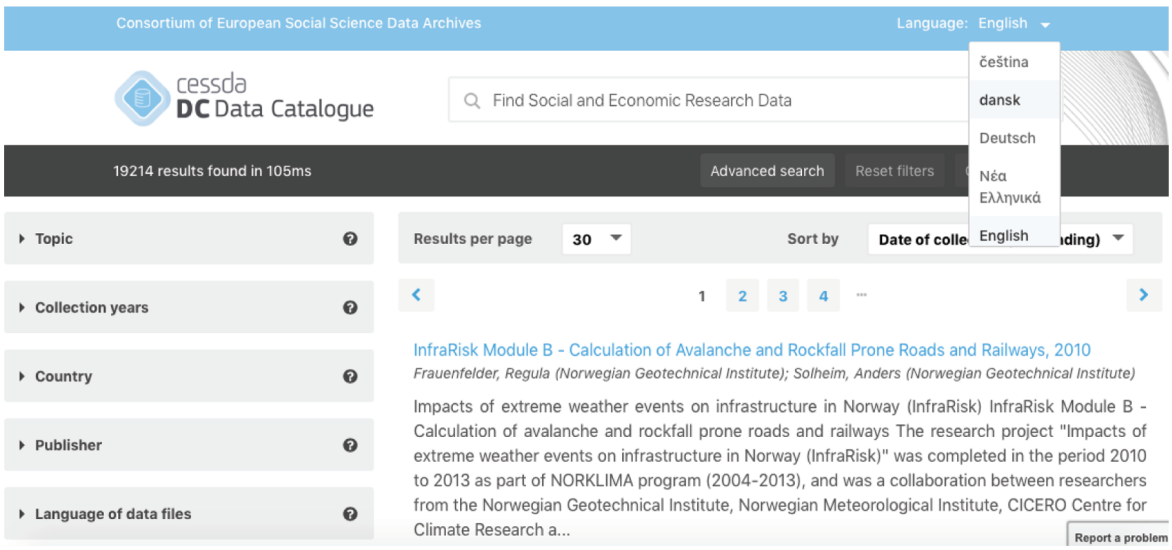

Figure 1. Screenshot from CESSDA Data Catalogue.

To produce a catalogue, CESSDA had to build a metadata harvester – collecting the metadata from the Service Providers. As these providers use different systems, we had to set up different end-points. Another tool was the metadata profile: we had to agree on the obligatory fields, based on the DDI-standard[4], and

[4] http://www.ddialliance.org.

ensure compatibility with other standards (e.g. schema.org, OpenAIRE, DataCite). For the filtering (closed selection of topics) we had to set the Vocabularies and we needed a multilingual thesaurus to support the searching. The technical and architecture specifications allow machine-reading of the catalogue.

Still in the pipeline (Table 3) are search facilities for questions and to search within variable names and value labels. And in the planning, is to set up a secured environment for access and analysis of data, including a single sign on: logging in on CESSDA gives access to secured environments of the national data service providers. Not in the table, but still an issue: the use of persistent identifiers (see also [7] and [8]). CESSDA has developed a Persistent Identifier Policy, that is in the pipeline for implementation.

Table 3. Elements of a Data Catalogue.

Tools	Purpose
CESSDA Data Catalogue	
CESSDA Metadata Harvester	Collect metadata (daily)
CESSDA Metadata Profile	Obligatory fields for describing datasets
Common Vocabularies	Needed for the filters
Multilingual Thesaurus	For (free) searching
	Under construction
European Question Bank	Search among survey questions
Variables & Values Metadata	Search among variables and value labels
	To Do
Single Sign On	Users sign in on secured CESSDA platform
Secured Access	Tool for access to sensitive/secured data

2.2.1 Conclusion

Building a catalogue relies on a good plan – on architecture, on pipelines to collect content in an automated way, on machine-readability. The catalogue is still under development, adding new functionality and new search options. But we think it was a good decision to go live with a basic version – to do further testing on real metadata and endpoints – and learn from users' feedback.

2.3 Strategy on FAIR – also work on Accessibility, Interoperability and Reuse

Access to social data can be very complicated, because of the security issues. Within CESSDA we work on a common approach on how to describe the different access levels in a uniform way.

Interoperability has big opportunities in social sciences: traditional work is on making data comparable over countries – doing this ex post implies a lot of work, so ideally this is done beforehand, e.g. European

Social Survey and the Survey of Health, Aging and Retirement in Europe. New work makes use of linked open data techniques and would allow to connect related data, e.g. in Election Studies: newspapers, social network data, party manifestos, surveys on election behaviour and motivations, etc.

Reusability deals with connecting data, data provenance and quality, connecting with research communities. This follows from the other principles and deals with a final stage of putting all the principles together. Before we can do this, we need to have a better understanding of all the principles and have more experience in dealing with them (see also [9] for more information on developing FAIR tools).

3. NEXT STEPS

Using the FAIR Action Plan Recommendations, we could picture where CESSDA is in implementing FAIR Data Principles. CESSDA follows a sequential approach, currently realising Findability, and working on the other principles.

Next steps will be on implementation of new tools, but also on cooperation – within the CESSDA consortium, within Social Sciences and Humanities – via the SSHOC project, and with other established research infrastructures – using the ERIC-Forum.

But we should look further, and seek cooperation with other information service organisations, to build a European Open Science Cloud that is beneficial to researchers of all disciplines and professionals and experts outside academia.

REFERENCES

[1] European Commission, Turning FAIR into reality, Final Report and Action Plan from the European Commission Expert Group on FAIR Data, Directorate-General for Research and Innovation, 2018. doi:10.2777/1524.

[2] ESFRI, European Strategy Forum on Research Infrastructures, Strategy Report on Research Infrastructures, Roadmap 2018, Directorate-General for Research and Innovation, 2018. isbn: 978-88-943243-3-4.

[3] P. Wittenburg, F. de Jong, D. van Uytvanck, M. Cocco, K. Jeffery, M. Lautenschlager, ... & P. Holub. State of FAIRness in ESFRI projects. Data Intelligence 2(2020), 230–237. doi: 10.1162/dint_a_00045.

[4] European Commission, LAB – FAB – APP, investing in the European future we want, Report of the independent High Level Group on maximising the impact of EU Research & Innovation Programmes, Directorate-General for Research and Innovation, 2017. doi:10.2777/30011.

[5] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, ... & E. Schultes. FAIR principles: Interpretations and implementation considerations. Data Intelligence 2(2020), 10–29. doi: 10.1162/dint_r_00024.

[6] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak ... & B. Mons. The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3(2016), 160018. doi: 10.1038/sdata.2016.18.

[7] T. Weigel, U. Schwardmann, J. Klump, S. Bendoukha & R. Quick. Making data and workflows findable for machines. Data Intelligence 2(2020), 40–46. doi: 10.1162/dint_a_00026.

[8] P. Groth, H. Cousijn, T. Clark & C. Goble. FAIR data reuse – the path through data citation. Data Intelligence 2(2020), 78–86. doi: 10.1162/dint_a_00030.

[9] M. Thompson, K. Burger, R. Kaliyaperumal, M. Roos & L.O. Bonino da Silva Santos. Making FAIR easy with FAIR tools: From creolization to convergence. Data Intelligence 2(2020), 87–95. doi: 10.1162/dint_a_00031.