Data integration with biological databases: gathering of technology

Alexander Vasilenko , Oleg Stupar, Galina Kochkina, Svetlana Ozerskaya

FAIR in the nearest systems

1. The Action Plan of EC Expert Group on Turning FAIR data into reality
2. WP6 in EOSC-Life
3. MIRRI

Key principle: FAIR (FORCE11)

Findable

F1. (Meta)data are assigned a globally unique and persistent identifier
F2. Data are described with rich metadata
F3. Metadata clearly and explicitly include the identifier of the data they describe
F4. (Meta)data are registered or indexed in a searchable resource

Accessible

A1. (Meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 The protocol is open, free, and universally implementable
A1.2 The protocol allows for an authentication and authorization procedure, where necessary
A2. Metadata are accessible, even when the data are no longer available

Interoperable

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (Meta)data use vocabularies that follow FAIR principles
I3. (Meta)data include qualified references to other (meta)data

Reusable

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (Meta)data are released with a clear and accessible data usage license
R1.2. (Meta)data are associated with detailed provenance
R1.3. (Meta)data meet domain-relevant community standards

Degrees of FAIR: a five star scale [1]

*	The basic core: metadata, PID & access	F2. data are described with rich metadata F1. (meta)data are assigned a globally unique and persistent identifier A1. (meta)data are retrievable by their identifier using a standardized communications protocol
**	Enhanced access: catalogues for discovery, standard (controlled) access & licences	F4. (meta)data are registered or indexed in a searchable resource A1.1. the protocol is free, open and universally implementable A1.2. the protocol allows for an authentication and authorization procedure, where necessary R1.1. (meta)data are released with a clear and accessible data usage license
***	Use of standards: for metadata and data	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation R1.3. (meta)data meet domain relevant community standards F3. metadata clearly and explicitly include the identifier of the data it describes
****	Rich, FAIR metadata	R1. (meta)data are richly described with a plurality of accurate and relevant attributes I2. (meta)data uses vocabularies that follow FAIR principles
*****	Provenance and additional context	R1.2 (meta)data are associated with data provenance I3. (meta)data include qualified references to other (meta)data A2. metadata are accessible, even when the data are no longer available

[1] European Commission Expert Group on Turning FAIR data into reality Action Plan Interim recommendations

Turning FAIR data into reality
Action Plan

Step 1: Define and apply FAIR appropriately

Rec. 1: Definitions of FAIR

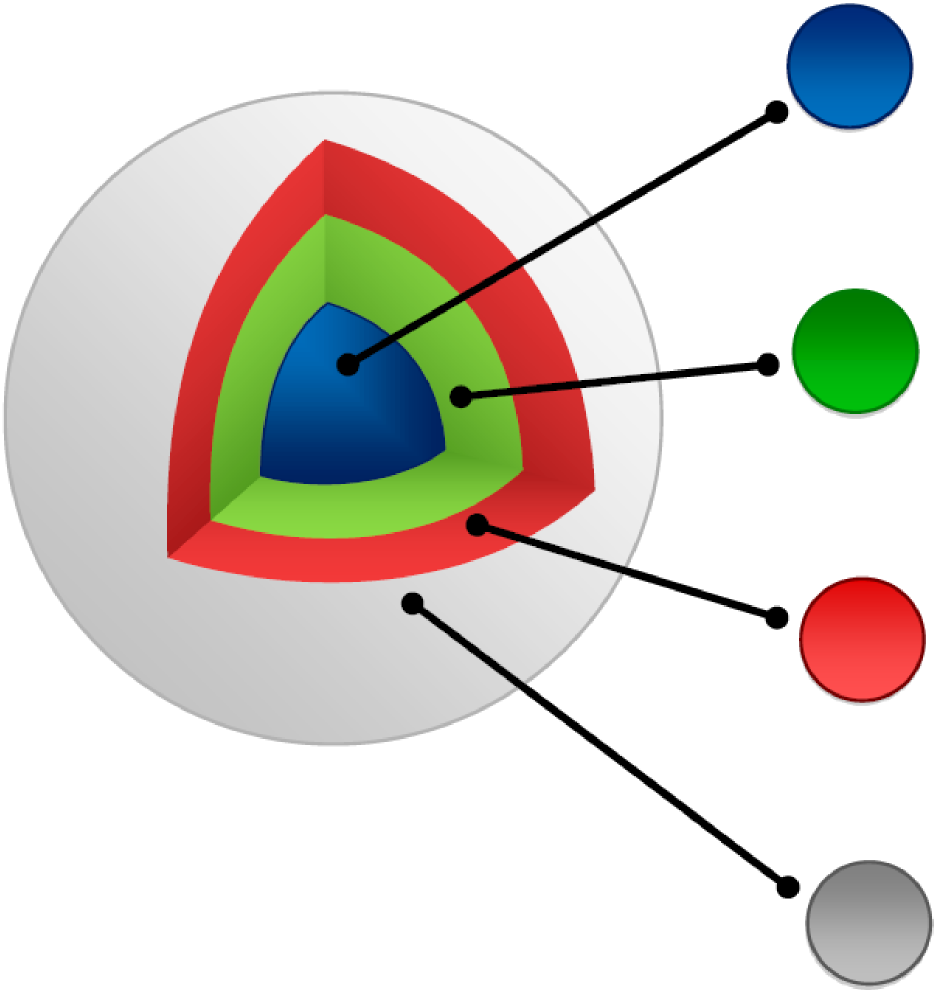
FAIR is not limited to its four constituent elements: it must also comprise appropriate openness, the assessability of data, long-term stewardship, and other relevant features. To make FAIR data a reality, it is necessary to incorporate these concepts into the definition of FAIR. (see also Rec. 2, 7)

Rec. 2: Mandates and boundaries for Open: «as open as possible, as closed as necessary» (see also Rec. 1)

Rec. 3: A model for FAIR Data Objects

Implementing FAIR requires a model for FAIR Data Objects which by definition have a PID linked to different types of essential metadata, including provenance and licensing. The use of community standards and sharing of code is also fundamental for interoperability and reuse.

General datamodel



DATA

The core bits

At its most basic level, data is a bitstream or binary sequence. For data to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and code. These layers of meaning enrich the data and enable reuse.

IDENTIFIERS

Persistent and unique (PIDs)

Data should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE

Open, documented formats

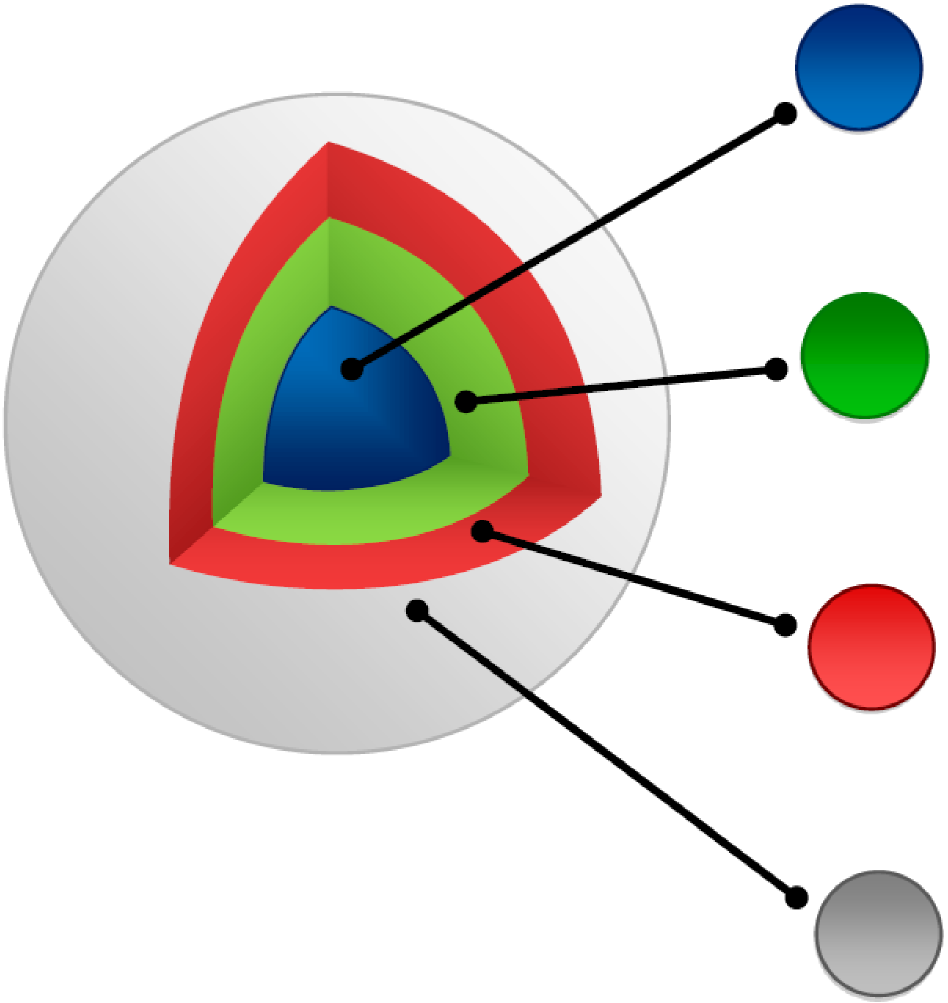
Data should be represented in common and ideally open file formats. This enables others to reuse the data as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.

METADATA

Contextual documentation

In order for data to be assessable and reusable, it should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the data were created. To enable the broadest reuse, data should be accompanied by a 'plurality of relevant attributes' and a clear and accessible data usage license.

mCC datamodel



DATA

The core bits

Strain passport

IDENTIFIERS

Persistent and unique (PIDs)

Strain number PID

STANDARDS & CODE

Open, documented formats

OECD MDS

METADATA

Contextual documentation

MTA, MAA, MDA, PIC, taxonomy used, mCC, curator, storage method, supply method, price, provenance, licensing

Action Plan

Step 2: Develop and support a sustainable FAIR data ecosystem

Rec. 4: **Components of a FAIR data ecosystem**

The realisation of FAIR data relies on, at minimum, the following essential components: policies, DMPs, identifiers, standards and repositories. There need to be registries cataloguing each component of the ecosystem and automated workflows between them. (see also Rec. 5, 25)

Rec. 5: **Sustainable funding for FAIR components**

The components of the FAIR ecosystem need to be maintained at a professional service level with sustainable funding. (see also Rec. 33, 11)

Rec. 6: **Strategic and evidence-based funding**

Funders of research data services should consolidate and build on existing investments in infrastructure and tools, where they demonstrate impact and community adoption. Funding should be tied to certification schemes as they develop for each of the FAIR ecosystem components. (see also Rec. 23, 34)

Action plan

Step 3: Ensure FAIR data and certified services

Rec. 7: **Disciplinary interoperability frameworks**

Research communities must be supported to develop and maintain their disciplinary interoperability frameworks. These incorporate principles and practices for data management and sharing, community agreements, data formats, metadata standards, tools and data infrastructure. (see also Rec. 8, 16)

Rec. 8: **Cross-disciplinary FAIRness**

Interoperability frameworks should be articulated in common ways and adopt global standards where possible to enable interdisciplinary research. Common standards, intelligent crosswalks, brokering mechanisms and machine-learning should all be explored to break down silos. (see also Rec. 7)

Rec. 9: **Develop robust FAIR data metrics**

A set of metrics for FAIR Data Objects should be developed and implemented, starting from the basic common core of descriptive metadata, PIDs and access. The design of these metrics needs to be mindful of unintended consequences and they should be regularly reviewed and updated. (see also Rec. 11)

Rec. 10: **Trusted Digital Repositories**

Repositories need to be encouraged and supported to achieve CoreTrustSeal certification. The development of rival repository accreditation schemes, based solely on the FAIR principles, should be discouraged. (see also Rec. 11, 18)

Rec. 11: **Develop metrics to assess and certify data services**

Certification schemes are needed to assess all components of the FAIR data ecosystem. Like CoreTrustSeal, these should address aspects of service management and sustainability, rather than being based solely on FAIR principles which are primarily articulated for data and objects. (see also Rec. 10, 9)

Action plan

Step 4: Embed a culture of FAIR in research practice

Rec. 12: **Data management via DMPs**

Any research project should include data management as a core element necessary for the delivery of its scientific objectives, addressing this in a Data Management Plan. The DMP should be regularly updated to provide a hub of information on the FAIR data objects. (see also Rec. 21, 32)

Rec. 13: **Professionalise data stewardship roles**

Steps need to be taken to develop two cohorts of professionals to support FAIR data: data scientists embedded in those research projects which need them, and data stewards who will ensure the management and curation of FAIR data. (see also Rec. 28, 14)

Rec. 14: **Recognise and reward FAIR data and data stewardship**

FAIR data should be recognized as a core research output and included in the assessment of research contributions and career progression. The provision of infrastructure and services that enable FAIR data must also be recognized and rewarded accordingly. (see also Rec. 13)

FAIR data policy

Rec. 15: **Policy harmonization**

Efforts should be made to align and consolidate FAIR data policy, reducing divergence, inconsistencies and contradictions

Rec. 16: **Broad application of FAIR**

FAIR should be applied broadly to all objects (including metadata, identifiers, software and DMPs) that are essential to the practice of research, and should inform metrics relating directly to these objects. (see also Rec. 7)

FAIR data culture (see also Rec. 7, 8, 12, 14)

Rec. 17: **Selection and prioritization of FAIR Data Objects**

Research communities and data stewards should better define which FAIR data objects are likely to have long-term value and implement processes to assist the appraisal and selection of outputs that will be retained in the long term and made FAIR.

Rec. 18: **Deposit in Trusted Digital Repositories**

Research data should be made available by means of Trusted Digital Repositories, and where possible in those with a mission and expertise to support a specific discipline or interdisciplinary research community. (see also Rec. 10)

Rec. 19: **Encourage and incentivize data reuse**

Funders should incentivize data reuse by promoting this in funding calls and requiring research communities to seek and build on existing data wherever possible.

Rec. 20: **Support legacy data to be made FAIR**

There are large amounts of legacy data that is not FAIR but would have considerable value if it were. Mechanisms should be explored to include some legacy data in the FAIR ecosystem where required.

Rec. 21: **Use information held in Data Management Plans**

DMPs hold valuable information on the data and related outputs, which should be structured in a way to enable reuse. Investment should be made in DMP tools that adopt common standards to enable information exchange across the FAIR data ecosystem.

Technology for FAIR (see also Rec. 3, 4, 10, 11)

Rec. 22: **Develop FAIR components to meet research needs**

While there is much existing infrastructure to build on, the further development and extension of FAIR components is required. These tools and services should fulfil the needs of data producers and users, and be easy to adopt. (see also Rec. 7, 8)

Rec. 23: **Incentivise services to support FAIR data**

Research facilities, in particular those of the ESFRI and national Roadmaps, should be incentivised to provide FAIR data by including it as a criteria in the initial and continuous evaluation process. Strategic research investments should consider service sustainability. (see also Rec. 5, 10, 11)

Rec. 24: Support semantic technologies

Semantic technologies are essential for interoperability and need to be developed, expanded and applied both within and across disciplines. (see also Rec. 4, 8)

Rec. 25: **Facilitate automated processing**

Automated processing should be supported and facilitated by FAIR components. This means that machines should be able to interact with each other through the system, as well as with other components of the system, at multiple levels and across disciplines. (see also Rec. 3, 8, 21)

Skills and roles for FAIR (see also Rec. 13)

Rec. 26: **Data science and stewardship skills**

Data skills of various types, as well as data management, data science and data stewardship competencies, need to be developed and embedded at all stages and with all participants in the research endeavour.

Rec. 27: **Skills transfer schemes and brokering roles**

Skills transfer schemes should be supported to equip researchers from various domains with information management skills or vice versa. Such individuals will play an important role as intermediaries to broker relations between research communities and infrastructure services.

Rec. 28: **Curriculum frameworks and training**

A concerted effort should be made to coordinate, systematise and accelerate the pedagogy and availability of training for data skills, data science and data stewardship. (see also Rec. 13)

FAIR metrics (see also Rec. 9,11)

Rec. 29: **Implement FAIR metrics**

Agreed sets of metrics should be implemented and monitored to track changes in the FAIRness of datasets or data-related resources over time. (see also Rec. 9, 11, 14)

Rec. 30: **Monitor FAIR**

Funders should report annually on the outcomes of their investments in terms of FAIR and track how the landscape matures. Specifically, how FAIR are the research objects that have been produced and to what extent are the funded infrastructures certified and supportive of FAIR data.

Rec. 31: **Support data citation and next generation metrics**

Systems providing citation metrics for FAIR Data Objects and other research outputs should be provided. In parallel, next generation metrics that reinforce and enrich citation-centric metrics for evaluation should be developed. (see also Rec. 14, 19)

Costs and investment in FAIR (see also Rec. 5, 6)

Rec. **32**: **Costing data management**

Research funders should require data management costs to be considered and included in grant applications, where relevant. To support this, detailed guidelines and worked examples of eligible costs for FAIR data should be provided. (see also Rec. 12)

Rec. 33: **Sustainable business models**

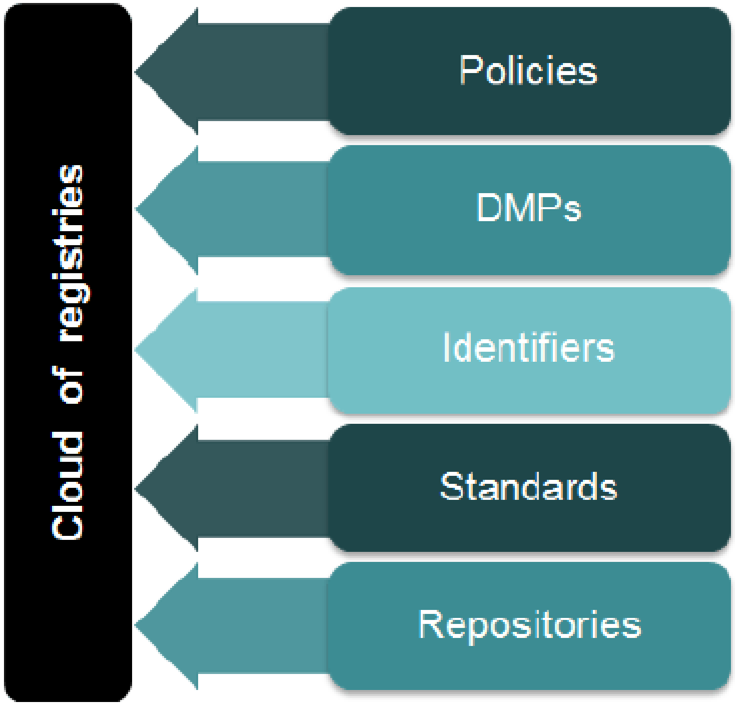
Data repositories and other components of the FAIR data ecosystem should be supported to explore business models for sustainability, to articulate their value proposition, and to trial a range of charging models and income streams. (see also Rec. 5, 32)

Rec. 34: **Leverage existing data services for EOSC**

The Rules of Engagement for EOSC must be broadly-defined and open to enable all existing service providers to address the criteria and be part of the European network. (see also Rec. 6)

Conclusions

FAIR SYSTEMS TO MAKE



.

FAIR ACTION PLAN RECS PARTIALLY PRESENTED IN MIRRI PLANS

1, 3, 7, 18, 20, 28, 34

FAIR ACTION PLAN RECS NOT PRESENTED IN MIRRI AT ALL

2, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 21, 22, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33

Acknowledgement

- WFCC team, and its former president Dr. Philippe Desmeth
- ECCO team, and its current president Prof. Nelson Lima
- WDCM team, and its director Dr. Juncai Ma, team leader Linhuan Wu
- Former StrainInfo team, and its leader Prof. Peter Dawyndt
- Former MIRRI WP8 team, and its leader Frank Oliver Glöckner

Thank you