

Towards a Common Standard for Data and Specimen Provenance in Life Sciences

Petr Holub^{*1}, Rudolf Wittner^{1,2}, Cecilia Mascia³, Francesca Frexia³, Heimo Müller⁴, Markus Plass⁴, Clare Allocca⁵, Fay Betsou⁶, Tony Burdett⁷, Ibon Cancio⁸, Adriane Chapman⁹, Martin Chapman¹⁰, Mélanie Courtot⁷, Vasa Curcin¹⁰, Johann Eder¹¹, Mark Elliot¹², Katrina Exter¹³, Elliot Fairweather¹⁰, Carole Goble¹⁴, Martin Golebiewski¹⁵, Bron Kisler¹⁶, Andreas Kremer¹⁷, Sheng Lin-Gibson¹⁸, Anna Marsano¹⁹, Marco Mattavelli²⁰, Josh Moore²¹, Hiroki Nakae²², Isabelle Perseil²³, Ayat Salman^{24,25}, James Sluka²⁶, Stian Soiland-Reyes^{14,27}, Caterina Strambio-De-Castillia²⁸, Michael Sussman²⁹, Jason R. Swedlow²¹, Kurt Zatloukal⁴, and Jörg Geiger³⁰

¹BBMRI-ERIC, Graz, AT

²Institute of Computer Science & Faculty of Informatics, Masaryk University, Brno, CZ

³CRS4 – Center for Advanced Studies, Research and Development in Sardinia, IT

⁴Medical University Graz, AT

⁵National Institute of Standards and Technology, Gaithersburg, MD, USA

⁶LNS - Laboratoire national de santé, LU

⁷EMBL's European Bioinformatics Institute (EMBL-EBI), UK

⁸Plentzia Marine Station (PiE-UPV/EHU), University of the Basque Country, EMBRC-Spain

⁹University of Southampton, UK

¹⁰King's College London, UK

¹¹University of Klagenfurt, AT

¹²Department of Social Statistics, School of Social Sciences, The University of Manchester, UK

¹³Flanders Marine Institute (VLIZ), EMBRC-Belgium

¹⁴Department of Computer Science, The University of Manchester, UK

¹⁵Heidelberg Institute for Theoretical Studies (HITS gGmbH), Heidelberg, DE

¹⁶Independent consultant

¹⁷ITTM S.A., LU

¹⁸Biosystems and Biomaterials Division, NIST, USA

¹⁹Department of Biomedicine, The University of Basel, CH

²⁰SCI-STI-MM, École Polytechnique Fédérale de Lausanne, Lausanne, CH

²¹Centre for Gene Regulation and Expression and Division of Computational Biology, School of Life Sciences, University of Dundee, Dundee, UK

²²Japan bio-Measurement and Analysis Consortium, JPN

²³INSERM - Institut National de la Santé et de la Recherche Médicale, FR

²⁴Standards Council of Canada

²⁵Canadian Primary Care Sentinel Surveillance Network (CPCSSN) Department of Family Medicine, Queen's University, CA

²⁶Biocomplexity Institute, Indiana University, USA

41
42
43
44

²⁷Informatics Institute, University of Amsterdam, NL

²⁸Program in Molecular Medicine, UMass Medical School, USA

²⁹US Department of Agriculture, USA

³⁰Interdisciplinary Bank of Biomaterials and Data Würzburg (ibd_w), Würzburg, DE

45 The profound crisis of scientific reproducibility has its roots in the enhanced avail-
46 ability of large volumes of data that are produced at ever increasing velocity, which
47 in turn often leads to the dissolution of the control mechanisms that traditionally en-
48 sured the quality of data and processes [1–7]. At the same time the origin and history
49 of specimens used to generate research data often remains inexplicit. While consid-
50 erable effort has been put in the development of standards for specimen quality, the
51 actual documentation has been left to the discretion of the provider of the specimen.
52 As a result the situation is exacerbated by the lack of consistent and comprehensive
53 documentation of specimens, which could support the identification of suspected, or
54 proven use of, fabricated data or specimen of unclear origin. Hence, the urgent need
55 for the trustworthy documentation of the data lineage and specimens is evident, espe-
56 cially when considering the serious impact of irreproducible or even flawed scientific
57 results on health, economics, and political decisions [8–12].

58 It is generally accepted that the properties and quality attributes of specimens
59 used in the life sciences have significant impact on the reliability of data generated
60 in downstream analytical procedures [13–15]. Experts from multiple life sciences do-
61 mains have called for the improvement and standardization of the documentation
62 of research and scientific service processes [16–22]. This has led in turn to the pro-
63 gressive development and implementation of data management and other functional
64 tools, such as discovery services, access pipelines, and standardized data models, en-
65 abling the sharing of data and specimens [23–28]. In practice, however, there remains
66 a gap between the needs and the reality of the requirements specified in accepted
67 standards, including technical, operational and legal specifications needed to ensure
68 the trustworthiness and traceability of data and specimens. Electronic lab notebooks
69 (ELN) and laboratory information management systems (LIMS) adopted by research
70 organizations might be considered attempts to electronically manage research work-
71 flows and data to promote reproducibility and traceability. However, these systems
72 can not provide the degree of standardization an international standard would offer,
73 as they are often proprietary and not subject to certification. In an effort to remedy
74 these deficiencies in the provenance captured and reported, we are endeavoring to de-
75 velop an *international standard on provenance information system for the life sciences*
76 accepted by both academia and industry. Provenance information can be used to as-
77 sess the quality and reliability, and hence the reusability of the object, i.e. the data,
78 the metadata, the biological materials, or the specimens.

79 The need for an effort to address the issues in provenance was proposed to the In-
80 ternational Standards Organization (ISO) Technical Committee 276 “Biotechnology”
81 (ISO/TC 276) in 2017 and approved as a preliminary work item. In 2020, ISO/TC 276

*Corresponding author: Assoc. Prof. RNDr. Petr Holub, Ph.D., Neue Stiftingtalstr. 2/B/6, 8010 Graz, AT.
Email: <petr.holub@bbmri-eric.eu>

82 approved a new work item proposal to develop an international standard for biological
83 material and data provenance and registered it as a working draft (WD), ISO/WD
84 23494-1 *Provenance information model for biological specimen and data — Part 1: Gen-*
85 *eral requirements*. This standardization effort is in accordance with the FAIR princi-
86 ples, which provide high-level methodological recommendations, including guidance
87 on provenance.¹ As the FAIR principles themselves do not provide detailed instruc-
88 tions for the implementation of provenance standards and documentation, ISO/WD
89 23494 is intended for data provenance of biological samples and will be built on the
90 World Wide Web Consortium’s (W3C) PROV [29], a generic provenance informa-
91 tion standard that defines a general model, corresponding serializations² and other
92 supporting specifications to enable the interoperable exchange of provenance infor-
93 mation between data environments. W3C PROV serves as a framework that is adapt-
94 able and extensible to fit the needs of diverse domains. The W3C PROV standard
95 has already been adopted in life science research areas [30], e.g., for computational
96 workflows [31], pharmacologic pipelines [32], neuroscience [33, 34], microscopy ex-
97 periments [35], medical sciences [36] and health implementation care³ in HL7 FHIR
98 [37]. Unfortunately, these implementations occurred without coordination and the
99 resulting solutions are often incompatible, incomplete, expressed at different levels of
100 granularity, and do not use a consistent approach for creating a continuous chain of
101 provenance from the “source” to the resulting data. Instead of redefining the W3C
102 PROV concepts, we have identified gaps that need to be filled in order to develop a
103 distributed, fully technically and semantically interoperable provenance information
104 standard that covers a given specimen and its associated metadata, and describes its
105 uninterrupted history from its “source”. The “source” can include a complex, multi-
106 institutional environment and can be both the source specimen and data, but also
107 link to a specific biological entity, or environmental specimen collected at a given
108 time and location (*connectivity* requirement [38]). The main goals of the provenance
109 information standard are

- 110 (i) to support improved reproducibility of life-sciences research, to provide a
111 voluntary provenance framework enabling concordance of governments, busi-
112 nesses, academia and the international community
- 113 (ii) to achieve harmonization of documentation of specimens that is compliant
114 with international conventions, recognized ethical practices and legal require-
115 ments such as the Nagoya Protocol [39] and the Declaration of Taipei [40].
- 116 (iii) to enable decision-making about the fitness-for-purpose of particular spec-
117 imens and data, by collecting and linking provenance information from the
118 whole life-cycle of the object (from specimen collection and processing, through
119 data generation and analysis) as depicted in Figure 1.

120 The standard will enhance the trustworthiness of provenance information by includ-
121 ing requirements and guidelines on its integrity, authenticity, and non-repudiation
122 [41], to prevent the production and/or use of unreliable, flawed or fabricated data

¹ Principle R1.2: (Meta)data are associated with detailed provenance.

² As defined in ISO 21597-1:2020: encoding of an ontology or dataset into a format that can be stored, typically in a file.

³ <https://www.hl7.org/fhir/provenance.html>

123 (the potential harms of which have become evident during the COVID-19 pandemic
124 [2, 10]), as well as accidental or malicious modification of data. Since provenance
125 information may also include sensitive or personal data (related, e.g., to the health
126 condition of an individual), the standard aims to enable sensitive information to be
127 concealed and disclosed only under strictly controlled conditions, while preserving its
128 core properties of integrity, authenticity and non-repudiation. Additional advanced
129 application scenarios include tracking of provenance information to: (i) track research
130 error propagation, (ii) identify people affected by incidental research findings, (iii)
131 check compliance with applicable regulations, or (iv) support production of reference
132 material by maintaining full documentation of provenance (complementing work of
133 ISO/TC 334 [42]). For research concerned with highly regulated fields in life sci-
134 ences, such as development of medical products or drugs, the standardized proven-
135 ance model will also contribute to a level of accountability and auditability of re-
136 search organisations.

137 The proposed standard is designed to cover the majority of the organizations in-
138 volved in life-sciences research, both academic and industrial, government labs and
139 research centers. Included organizations are university and industrial research labo-
140 ratories, biobanks and biorepositories, culture collections, hospitals, research centres,
141 and private companies (e.g., pharmaceutical companies or lab reagent suppliers). The
142 broader audience includes not only research data producers, but also those publishing,
143 cataloguing, archiving or reusing research data [43]. The standard can also be adopted
144 by manufacturers and vendors of laboratory instruments – e.g., automation devices,
145 microscopes, sequencers, spectrometers – to enable automated standard-compliant
146 generation of provenance information. Automated generation of provenance infor-
147 mation will minimize human errors and the burden put on workers, both in terms
148 of effort and training. Provenance information generated automatically by devices
149 should be interoperable to enable automated integration and quality control as well
150 as validity checks demonstrating standard-compliant provenance. The standard is in-
151 tended to cover a wide range of research and applications in life sciences and for that
152 reason a modular structure has been used to enable extensibility to evolving require-
153 ments, processes, or technologies.

154 The current draft proposal ISO/WD TS 23494 1 is the first part of a planned series
155 of six parts, with the intent that each will become a distinct ISO standard:

- 156 1. *Provenance Information Management* defines the overall structure of the stan-
157 dard and provides general requirements on provenance information manage-
158 ment, thus enabling interconnections between the various components of prove-
159 nance information in distributed environments. It also specifies requirements
160 applicable to entities responsible for generating the provenance information.
- 161 2. The *Common Provenance Model* builds on the W3C PROV model, defining repre-
162 sentations of elements common to all stages of research, such as interlinking of
163 distributed components of provenance information by sender and receiver ob-
164 jects, the identification of physical and digital objects, and provisions for non-
165 repudiation. Provenance information patterns for common scenarios, such as
166 the compound processes, versioning of provenance information or documen-

167 tation of accountabilities. The model will also define mechanisms to embed or
168 reference entire records of provenance information.

169 3. *Provenance of Biological Materials* defines requirements and scope of the prove-
170 nance information documenting biological material or specimen acquisition,
171 handling and processing and builds on the Common Provenance Model. This
172 includes, but is not limited to, data on collection and collection procedure, trans-
173 port conditions, and documentation of legal and ethical basis (e.g. consent,
174 terms of access and benefit sharing). It will also provide mechanisms to refer-
175 ence Standard Operating Procedures (SOPs) and compliance with or deviations
176 from them. Referencing the widely accepted de-facto reporting standard for bi-
177 ological specimen quality SPREC [44] will also be enabled. Actual techniques
178 or practices for handling biological material are not specified in the standard, in
179 favor of technical specifications enabling consistent interoperable and machine-
180 actionable documentation of handling biological material. With the provenance
181 information provided, however, the standard facilitates the verification of com-
182 pliance with other pre-analytical ISO standards covering biobanking, analyti-
183 cal and processing methods, generation of reference material and related fields
184 (ISO 20387:2018, ISO 20184 series, ISO 20166 series, and ISO 20186 series).

185 4. *Provenance of Data Generation* defines the provenance of data generated from
186 the analysis or observation of biological material, e.g., sequencing, microscopy,
187 spectrometry, etc. Provenance information specific for diverse analytical or ob-
188 servational methods will be embedded in a way meeting the requirements of the
189 particular domain, but as well compliant with the provenance model standard
190 allowing seamless integration in a complete provenance chain.

191 5. *Provenance of Data Processing* defines provenance of computational aspects of
192 life sciences research (such as computational workflows based on CWLProv [31]
193 and RO Crate [45]).

194 6. *Security Extensions* define optional extensions supporting authenticity, integrity
195 and non-repudiation of provenance information, and hence its trustworthiness
196 and reliability. Demonstration of these properties will also be supported for
197 sensitive elements of provenance information.

198 The ISO standards development process responds to a market need and is based
199 on globally-relevant expertise. The product is a voluntary consensus standard de-
200 veloped through a multi-stakeholder process. ISO/WD 23494-1 has a proven market
201 need and has passed through the preliminary stages of the ISO voting process – as
202 a result, it is part of the ISO Work Programme. The document is under development
203 and will evolve along the multi-stage ISO standard development process. ISO/WD
204 23494-1 *Provenance information model for biological specimen and data – Part 1: Gen-
205 eral requirements* is currently at the working draft stage, and is anticipated to move
206 next to the committee draft (CD) stage. The document will be revised and reviewed
207 throughout the ISO project stages until final approval and publication. Part 2 of this
208 series, *Biotechnology – Provenance information model for biological material and data*

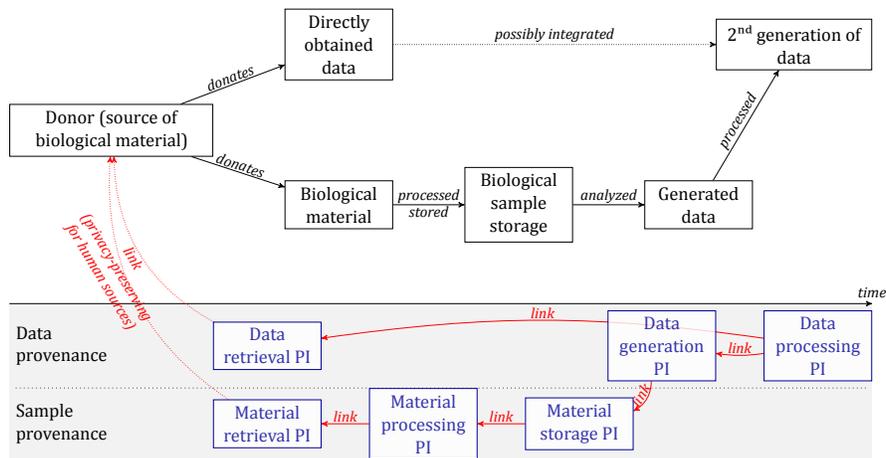


Figure 1: An overview of provenance chain. A sample obtained from a donor (or other source) is created and an initial set of provenance information (PI) is generated. As that sample moves through time and space, is processed and/or analyzed, additional provenance data is appended to the provenance chain for each new item. The chain can be extended as a complete unit of later stages of provenance or use unique identifiers to refer to early stages of provenance data.

209 — Part 2: Common provenance model, has been accepted by ISO/TC 276/WG 5 as pre-
 210 liminary work item ISO/PWI TS 23494-2 . The future documents in this series are in
 211 planning stages, but not yet submitted to ISO/TC 276/WG 5 . The standards develop-
 212 ment process builds on existing standards for collection and processing of specimens,
 213 analytical techniques and data generation and analysis, as well as use-cases from the
 214 biomedical domain. BBMRI-ERIC, which is also active in developing international
 215 standards for biobanking, has drafted use-cases for biological material provenance.
 216 Collaborations and ISO liasions with professional societies like the European, Middle
 217 Eastern and African Society for Biobanking (ESBB) and the International Society for
 218 Biological and Environmental Repositories (ISBER) have also contributed to the devel-
 219 opment of specimen provenance use cases. In addition, use cases on data generation
 220 and processing can come from subject matter experts and the scientific community
 221 including the European EOSC-Life project,⁴ Open Microscopy Environment, OME,⁵
 222 genetic data compression (ISO/IEC JTC1/SC 29/WG 08 MPEG-G) [46], clinical deci-
 223 sion support systems (Kings College London) and other life sciences domains such as
 224 biodiversity, marine biology and systems biology.

225 However, alternatives to ISO standards process⁶ exist—some community-based ef-
 226 forts have developed widely adopted specifications that have become *de facto* global

⁴ <https://www.eosc-life.eu/>

⁵ <https://www.openmicroscopy.org/>

⁶ <https://www.iso.org/developing-standards.html>

227 standards.⁷ The success of these examples lies, at least in part, in the pairing of a
228 specification with an accessible implementation that validates the utility of the spec-
229 ification and allows a broad community to explore integration into applications that
230 extend far beyond the initial target [50]. We believe that community-led and ISO-
231 based approaches for developing and delivering standards can complement each other
232 and that a combination of parallel efforts for developing a provenance chain standard
233 might ultimately be the most productive approach. As the provenance information
234 model development is grounded in the EOSC-Life project, collaboration with these
235 communities is already established. The ISO standard development is thus considered
236 as a standardized instance of a publicly available model developed in parallel under
237 auspices of EOSC-Life [51].

238 Another challenge is the continuous dissemination and periodic revision of the
239 standard once published. Though ISO standards are not “open access”, they can be
240 purchased for a moderate fee⁸ or accessed through institutional libraries, and, bar-
241 ring any patent restrictions, can be freely implemented, for instance, in Open Source
242 software. ISO standards can also include Open Source reference implementations as
243 specific normative or informative parts of the standards. ISO standards can be im-
244 plemented independently or based on such source code, in compliance with the rea-
245 sonable and non-discriminatory (RAND) licensing terms imposed by the ISO require-
246 ments. Such licensing terms, like for instance the one applied to all ISO/IEC/SC29
247 (MPEG) standards that are free from any charge for scientific and non-profit research
248 purposes, may or may not include licensing fees.

249 We would like to invite experts from biotechnology and biomedical fields to fur-
250 ther contribute to the standard, in particular to the provenance of biological speci-
251 mens, the data-generation and data-processing modules. Help is needed to develop
252 applications of the general modules and the development of specific use cases, as well
253 as direct contributions to the text of the standard itself. Contributions are possible
254 through a liaison organization, a national ISO body or by engaging with EOSC-Life
255 project events and calls.

256 **Acknowledgments** This work has been co/funded by EOSC-Life supported by EU
257 Horizon 2020, grant agreement no. 824087; EJP-RD supported by EU Horizon 2020,
258 grant agreement no. 825575; BioExcel-2 supported by EU Horizon 2020, grant agree-
259 ment no. 823830; the DIFRA Project, funded by the Sardinian Regional Authority.
260 VC, EF, and MCh supported by the National Institute for Health Research (NIHR)
261 Biomedical Research Centre based at Guy’s and St Thomas’ National Health Service
262 NHS) Foundation Trust and King’s College London. TB, MCo acknowledge funding
263 from EMBL-EBI Core Funds and the FAIRplus project (H2020 No 802750). MCo sup-
264 ported by Wellcome Trust GA4GH award number 201535/Z/16/Z and the CINECA
265 project (H2020 No 825775). AC was supported by EPSRC (EP/S028366/1). JS was sup-
266 ported by the US National institute of Health (U24 EB028887, R01 GM122424, and
267 OT2OD026671). ME was supported by the Alan Turing Institute (ProvAnon). The

⁷ E.g., for on-line cryptography (RSA public keys [47]), scientific workflows (Common Workflow Language [48]) and bioimaging data formats (OME-TIFF [49]).

⁸ In some cases ISO standards can be obtained without any fee, e.g. <https://www.iso.org/covid19>

268 opinions in this paper are those of the authors and do not necessarily reflect the opin-
269 ions of the funders.

270 **Representation of communities** The co-authors team represents wide coverage
271 of life-sciences communities. PH, RW, CM, FF, HM, MP, JG come from human biobank-
272 ing and biomolecular resources communities, BBMRI-ERIC Research Infrastructure,
273 and are directly involved as experts in the ISO standardization process. KZ and JE
274 come from cancer research, biobanking and medical informatics and are long-term
275 contributors to data quality standardization efforts. TB, MCo lead development of the
276 BioSamples database at EMBL-European Bioinformatics Institute. IC and KE come
277 from marine biology and EMBRC Research Infrastructure. CG and SS-R have worked
278 with bioinformatics, CWL, RO-Crate and the original W3C PROV standards develop-
279 ments. JRS and JM come from bio-imaging communities and EUBioImaging Research
280 Infrastructure. VC, EF, and MCh come from health informatics. HN participates in
281 provenance standardization process as an expert from Japan, MS and JS as experts
282 from the U.S.A, and AK as an expert from Luxembourg. ME contributes to privacy
283 protection and provenance aspects. FB is a biobanking expert and chairing the ISBER
284 Biospecimen Science Working Group. AS is a biobanking expert and ESBB council-
285 lor. SL-G and CA are from NIST and convenor and secretary of ISO/TC 276/WG 3
286 "Analytical Methods". AM belongs to the tissue engineering and biomedical research
287 community. MM is a standard expert in the digital media, genomic sequencing and an-
288 notation data fields, and convenor of ISO/IEC SC29/WG 8 "MPEG Genomic Coding".
289 AC contributes to capture and handling of provenance within large organizations.

References

- 291 1. Begley CG and Ioannidis JP. Reproducibility in Science. *Circulation Research*
292 2015;116:116–26. DOI: 10.1161/CIRCRESAHA.114.303819.
- 293 2. Servick K and Enserink M. The pandemic’s first major research scandal erupts.
294 *Science* 2020;368:1041–2. DOI: 10.1126/science.368.6495.1041.
- 295 3. Lagoze C. Big Data, data integrity, and the fracturing of the control zone. *Big*
296 *Data & Society* 2014;1:2053951714558281. DOI: 10.1177/2053951714558281.
- 297 4. Mobley A, Linder SK, Braeuer R, et al. A Survey on Data Reproducibility in Can-
298 cer Research Provides Insights into Our Limited Ability to Translate Findings
299 from the Laboratory to the Clinic. *PLOS ONE* 2013;8:1–4. DOI: 10.1371/journal.
300 pone.0063221.
- 301 5. Morrison SJ. Time to do something about reproducibility. *eLife* 2014;3:1–4. DOI:
302 10.7554/eLife.03981.
- 303 6. Byrne JA, Grima N, Capes-Davis A, et al. The Possibility of Systematic Research
304 Fraud Targeting Under-Studied Human Genes: Causes, Consequences, and Po-
305 tential Solutions. *Biomarker Insights* 2019;14. DOI: 10.1177/1177271919829162.
- 306 7. Prinz F, Schlange T, and Asadullah K. Believe it or not: how much can we rely
307 on published data on potential drug targets? *Nature Reviews Drug Discovery*
308 2011;10. Number: 9 Publisher: Nature Publishing Group:712–2. DOI: 10.1038/
309 nrd3439-c1.
- 310 8. Freedman LP, Cockburn IM, and Simcoe TS. The Economics of Reproducibility
311 in Preclinical Research. *PLOS Biology* 2015;13:1–9. DOI: 10.1371/journal.pbio.
312 1002165.
- 313 9. Nickerson D, Atalag K, Bono B de, et al. The Human Physiome: how standards,
314 software and innovative service infrastructures are providing the building blocks
315 to make it achievable. *Interface Focus* 2016;6. 00001:20150103. DOI: 10.1098/
316 rsfs.2015.0103. URL: [http://rsfs.royalsocietypublishing.org/lookup/doi/](http://rsfs.royalsocietypublishing.org/lookup/doi/10.1098/rsfs.2015.0103)
317 [10.1098/rsfs.2015.0103](http://rsfs.royalsocietypublishing.org/lookup/doi/10.1098/rsfs.2015.0103) (visited on 03/10/2016).
- 318 10. Mahase E. Covid-19: 146 researchers raise concerns over chloroquine study that
319 halted WHO trial. *BMJ* 2020;369. DOI: 10.1136/bmj.m2197.
- 320 11. Chaplin S. Research misconduct: how bad is it and what can be done? *Future*
321 *Prescriber* 2012;13:5–76. DOI: 10.1002/fps.88.
- 322 12. Committee on Responsible Science, Committee on Science, Engineering, Medicine,
323 and Public Policy, Policy and Global Affairs, et al. *Fostering Integrity in Re-*
324 *search*. Pages: 21896. Washington, D.C.: National Academies Press, 2017. DOI:
325 10.17226/21896.
- 326 13. Simeon-Dubach D and Perren A. Better provenance for biobank samples. *Nature*
327 2011;475:454–5. DOI: 10.1038/475454d.
- 328 14. Holub P, Kohlmayer F, Prasser F, et al. Enhancing Reuse of Data and Biological
329 Material in Medical Research: From FAIR to FAIR-Health. *Biopreservation and*
330 *Biobanking* 2018;16:97–105. DOI: 10.1089/bio.2017.0110.

- 331 15. Müller H, Reihls R, Zatloukal K, et al. State-of-the-Art and Future Challenges in
332 the Integration of Biobank Catalogues:13. DOI: 10.1007/978-3-319-16226-3_11.
- 333 16. Ioannidis JP, Greenland S, Hlatky MA, et al. Increasing value and reducing waste
334 in research design, conduct, and analysis. *The Lancet* 2014;383:166–75. DOI: 10.
335 1016/S0140-6736(13)62227-8.
- 336 17. Freedman LP and Inglese J. The Increasing Urgency for Standards in Basic Bio-
337 logic Research. *Cancer Research* 2014;74:4024–9. DOI: 10.1158/0008-5472.CAN-
338 14-0925.
- 339 18. Begley CG and Ellis LM. Drug development: Raise standards for preclinical cancer
340 research. *Nature* 2012;483:531–3. DOI: 10.1038/483531a. arXiv: 9907372v1.
- 341 19. Landis SC, Amara SG, Asadullah K, et al. A call for transparent reporting to opti-
342 mize the predictive value of preclinical research. *Nature* 2012;490. nature11556[PII]:187–
343 91. DOI: 10.1038/nature11556.
- 344 20. Consortium of European Taxonomic Facilities (CETAF) Code of Conduct and
345 Best Practice for Access and Benefit-Sharing. URL: [https://ec.europa.eu/
346 environment/nature/biodiversity/international/abs/pdf/CETAF%20Best%
347 20Practice%20-%20Annex%20to%20Commission%20Decision%20C\(2019\)%203380%
348 20final.pdf](https://ec.europa.eu/environment/nature/biodiversity/international/abs/pdf/CETAF%20Best%20Practice%20-%20Annex%20to%20Commission%20Decision%20C(2019)%203380%20final.pdf) (visited on 02/15/2021).
- 349 21. Benson EE, Harding K, and Mackenzie-dodds J. A new quality management per-
350 spective for biodiversity conservation and research: Investigating Biospecimen
351 Reporting for Improved Study Quality (BRISQ) and the Standard PRE-analytical
352 Code (SPREC) using Natural History Museum and culture collections as case
353 studies. *Systematics and Biodiversity* 2016;14:525–47. DOI: 10.1080/14772000.
354 2016.1201167.
- 355 22. A-E. K and Tillin H. The EMBRC guide to ABS compliance. Recommendations
356 to marine biological resources collections’ and users’ institutions. A handbook
357 produced by the European Marine Biological Resource Centre. European Ma-
358 rine Biological Resource Centre. 2020. URL: [https://bluebiobank.eu/docs/
359 EMBRCGuideABS.pdf](https://bluebiobank.eu/docs/EMBRGuideABS.pdf).
- 360 23. Villanueva AG, Cook-Deegan R, Koenig BA, et al. Characterizing the Biomedical
361 Data-Sharing Landscape. *The Journal of Law, Medicine & Ethics: A Journal of
362 the American Society of Law, Medicine & Ethics* 2019;47:21–30. DOI: 10.1177/
363 1073110519840481.
- 364 24. Hulsén T. Sharing Is Caring-Data Sharing Initiatives in Healthcare. *International
365 Journal of Environmental Research and Public Health* 2020;17:E3046. DOI: 10.
366 3390/ijerph17093046.
- 367 25. Banzi R, Canham S, Kuchinke W, et al. Evaluation of repositories for sharing
368 individual-participant data from clinical studies. *Trials* 2019;20:169. DOI: 10.1186/
369 s13063-019-3253-3.
- 370 26. Toh S. Analytic and Data Sharing Options in Real-World Multidatabase Studies
371 of Comparative Effectiveness and Safety of Medical Products. *Clinical Pharma-
372 cology and Therapeutics* 2020;107:834–42. DOI: 10.1002/cpt.1754.

- 373 27. Grossman RL. Data Lakes, Clouds, and Commons: A Review of Platforms for
374 Analyzing and Sharing Genomic Data. *Trends in genetics: TIG* 2019;35:223–34.
375 DOI: 10.1016/j.tig.2018.12.006.
- 376 28. Wilson SL, Way GP, Bittremieux W, et al. Sharing biological data: why, when,
377 and how. *FEBS Letters* 2021;595. eprint: <https://febs.onlinelibrary.wiley.com/doi/pdf/10.1002/1873-3468.14067:847-63>. DOI: 10.1002/1873-3468.14067. URL: <https://febs.onlinelibrary.wiley.com/doi/abs/10.1002/1873-3468.14067> (visited on 06/22/2021).
- 380 29. Groth P and Moreau L. PROV-Overview: An Overview of the PROV Family of
381 Documents. 2013. URL: <https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/>.
- 383 30. Huynh TD, Groth P, and Zednik S. PROV Implementation Report. 2013. URL:
384 <http://www.w3.org/TR/2013/NOTE-prov-implementations-20130430/>.
- 385 31. Khan FZ, Soiland-Reyes S, Sinnott RO, et al. Sharing interoperable workflow
386 provenance: A review of best practices and their practical application in CWL-
387 Prov. *GigaScience* 2019;8. giz095. DOI: 10.1093/gigascience/giz095.
- 388 32. Mammoliti A, Smirnov P, Safikhani Z, et al. Creating reproducible pharmaco-
389 genomic analysis pipelines. *Scientific Data* 2019;6:166. DOI: 10.1038/s41597-019-
390 0174-7.
- 391 33. McClatchey R, Shamdasani J, Branson A, et al. Traceability and Provenance in
392 Big Data Medical Systems. In: *2015 IEEE 28th International Symposium on Computer-
393 Based Medical Systems*. 2015:226–31. DOI: 10.1109/CBMS.2015.10.
- 394 34. Giesler A, Czekala M, Hagemeyer B, et al. UniProv: A Flexible Provenance Track-
395 ing System for UNICORE. In: *High-Performance Scientific Computing*. Ed. by Di
396 Napoli E, Hermanns MA, Iliev H, et al. Cham: Springer International Publishing,
397 2017:233–42. DOI: 10.1007/978-3-319-53862-4_20.
- 398 35. Samuel S. Integrative Data Management for Reproducibility of Microscopy Ex-
399 periments. In: *The Semantic Web*. Ed. by Blomqvist E, Maynard D, Gangemi A,
400 et al. Cham: Springer International Publishing, 2017:246–55. DOI: 10.1007/978-
401 3-319-58451-5_19.
- 402 36. Curcin V, Fairweather E, Danger R, et al. Templates as a method for implement-
403 ing data provenance in decision support systems. *Journal of Biomedical Informatics*
404 2017;65:1–21. DOI: 10.1016/j.jbi.2016.10.022.
- 405 37. HL7 and its participants. FHIR Release #4 [Standard], version 4.0.1. 2019. URL:
406 <http://hl7.org/fhir/R4/>.
- 407 38. Curcin V, Miles S, Danger R, et al. Implementing interoperable provenance in
408 biomedical research. *Future Generation Computer Systems* 2014;34. Special Sec-
409 tion: Distributed Solutions for Ubiquitous Computing and Ambient Intelligence:1–
410 16. DOI: 10.1016/j.future.2013.12.001.
- 411 39. Unit B. The Nagoya Protocol on Access and Benefit-sharing. Publisher: Secre-
412 tariat of the Convention on Biological Diversity. 2021. URL: <https://www.cbd.int/abs/> (visited on 06/22/2021).
- 413

- 414 40. WMA - The World Medical Association-WMA Declaration of Taipei on Ethical
415 Considerations regarding Health Databases and Biobanks. URL: <https://www.wma.net/policies-post/wma-declaration-of-taipei-on-ethical-considerations-regarding-health-databases-and-biobanks/> (visited on 07/02/2020).
- 418 41. Fairweather E, Wittner R, Chapman M, et al. Non-repudiable provenance for
419 clinical decision support systems. preprint. 2020. arXiv: 2006.11233 [cs.CR].
- 420 42. 14:00-17:00. ISO/WD Guide 85. URL: <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/55/75538.html> (visited on
421 06/09/2021).
- 423 43. Cheney J, Chapman AP, Davidson J, et al. Data provenance, curation and quality
424 in metrology. 2021. arXiv: arXiv:2102.08228v1.
- 425 44. Betsou F, Bilbao R, Case J, et al. Standard PREanalytical Code Version 3.0. Biop-
426 reservation and Biobanking 2018;16:9–12. DOI: 10.1089/bio.2017.0109. (Visited
427 on 02/18/2020).
- 428 45. Carragáin EÓ, Goble C, Sefton P, et al. A lightweight approach to research object
429 data packaging. In: *Bioinformatics Open Source Conference (BOSC) 2019*. 2019. DOI:
430 10.5281/zenodo.3250687.
- 431 46. Voges J, Hernaez M, Mattavelli M, et al. An Introduction to MPEG-G: The First
432 Open ISO/IEC Standard for the Compression and Exchange of Genomic Sequenc-
433 ing Data. Proceedings of the IEEE 2021. DOI: 10.1109/JPROC.2021.3082027.
- 434 47. Rivest RL, Shamir A, and Adleman L. A Method for Obtaining Digital Signatures
435 and Public-Key Cryptosystems. *Commun. ACM* 1978;21:120–6. DOI: 10.1145/
436 359340.359342. URL: <https://doi.org/10.1145/359340.359342>.
- 437 48. Crusoe MR, Abeln S, Iosup A, et al. Methods Included: Standardizing Computa-
438 tional Reuse and Portability with the Common Workflow Language. 2021. arXiv:
439 2105.07028v1 [cs.DC].
- 440 49. Linkert M, Rueden CT, Allan C, et al. Metadata matters: access to image data
441 in the real world. *Journal of Cell Biology* 2010;189:777–82. DOI: 10.1083/jcb.
442 201004104. URL: <https://doi.org/10.1083/jcb.201004104>.
- 443 50. Swedlow JR, Kankaanpää P, Sarkans U, et al. A global view of standards for open
444 image data formats and repositories. *Nature Methods* 2021. DOI: 10.1038/s41592-
445 021-01113-7. URL: <https://doi.org/10.1038/s41592-021-01113-7>.
- 446 51. Wittner R, Mascia C, Frexia F, et al. EOSC-Life Common Provenance Model.
447 EOSC-Life deliverable D6.2. 2021. DOI: 10.5281/zenodo.4705074.