ỦY BAN CHÂU ÂU (EC)
Dịch vụ EU PWC viết, tháng 3/2018

EUROPEAN COMMISSION
Written by PwC EU Services March – 2018

Chi phí của việc không có dữ liệu nghiên cứu FAIR

Phân tích chi phí - lợi ích đối với dữ liệu nghiên cứu FAIR

Dịch sang tiếng Việt: Lê Trung Nghĩa

Dịch xong: 02/04/2023

Bản gốc tiếng Anh: http://publications.europa.eu/resource/cellar/d375368c-1a0a-11e9-8d04-01aa75ed71a1.0001.01/DOC_1

Cost of not having FAIR research data

Cost-Benefit analysis for FAIR research data

Nghiên cứu và Đổi mới sáng tạo

Research and Innovation

Phân tích chi phí - lợi ích đối với dữ liệu nghiên cứu FAIR - chi phí của việc không có dữ liệu nghiên cứu FAIR

Ủy ban châu Âu

Ban Tổng giám đốc về Nghiên cứu và Đổi mới sáng tạo

Ban A - Phát triển và Điều phối Chính sách

Đơn vị A.2 - Chính sách Dữ liệu Mở và Đám mây Khoa học

Liên hệ Athanasios Karalopoulos

Thư điện tử Athanasios.Karalopoulos@ec.europa.eu

RTD-PUBLICATIONS@ec.europa.eu

Ủy ban châu Âu

B-1049 Brussels

Bản thảo hoàn thành vào tháng 3/2018.

Luxembourg: Văn phòng Xuất bản của Liên minh châu Âu, 2018

PDF ISBN978-92-79-98886-8 doi: 10.2777/02999 KI-02-19-023-EN-N

Chi phí của việc không có dữ liệu nghiên cứu FAIR

Phân tích chi phí - lợi ích đối với dữ liệu nghiên cứu FAIR

Dịch vụ EU của PwC viết

PwC

Mục lục

1. GIỚI THIỆU 9
1.1 Bối cảnh 9
1.1.1 Giá trị của FAIR 10
1.1.2 Các sáng kiến FAIR hiện hành 10
1.1.3 Các thách thức triển khai các nguyên tắc FAIR 12
1.2 Khán thính phòng được mong đợi 13
1.3 Dữ liệu FAIR so với dữ liệu mở 14
1.4 Cấu trúc của báo cáo 15
2. CÁCH TIẾP CẬN 16
2.1 Tổng quan 16
2.2 Xác định các chỉ số 17
2.2.1 Tác động lên các hoạt động nghiên cứu 19
2.2.2 Tác động lên cộng tác 25
2.2.3 Tác động lên đổi mới sáng tạo 27
2.3 Các chỉ số được lựa chọn 29
2.4 Định lượng các chỉ số 31
2.4.1 Chỉ số #1: Thời gian bỏ ra 32
2.4.2 Chỉ số #2: Chi phí lưu trữ 36
2.4.3 Chỉ số #3: Chi phí cấp phép 39
2.4.4 Chỉ số #4: Rút lại nghiên cứu 40
2.4.5 Chỉ số #5: Cấp vốn hai lần 41
2.4.6 Chỉ số #6: Liên ngành 42
2.4.7 Chỉ số #7: Tăng trưởng kinh tế tiềm năng 42

3. TÍNH TOÁN CHI PHÍ 45
3.1 Chỉ số #1: Thời gian bỏ ra 46
3.2 Chỉ số #2: Chi phí lưu trữ 47
3.3 Chỉ số #3: Chi phí cấp phép 47
3.4 Chỉ số #4: Rút lại nghiên cứu 48
3.5 Chỉ số #5: Cấp vốn hai lần 49
3.6 Chỉ số #6: Liên ngành 49
3.7 Chỉ số #7: Tăng trưởng kinh tế tiềm năng 50
4. KẾT LUẬN 52
TÀI LIỆU THAM KHẢO 55
PHỤ LỤC 1. CÁC NGUYÊN TẮC FAIR 61

Danh sách các hình ảnh

Hình 1: Bốn đặc tính cơ bản của FAIR 9
Hình 2: Trình bày mức cao phương pháp luận 16
Hình 3: Các hoạt động nghiên cứu 20
Hình 4: Phân bổ lượng các bản sao theo số lượng các kho 37
Hình 5: Phân bổ chi phí 45

Danh sách các bảng biểu

Bảng 1: Các chỉ số được lựa chọn cho từng lĩnh vực 29
Bảng 2: Tính toán chỉ số #1 46
Bảng 3: Tính toán chỉ số #2 47
Bảng 4: Tính toán chỉ số #3 47
Bảng 5: Tính toán chỉ số #4 48
Bảng 6: Tính toán chỉ số #5....49
Bảng 7: Các nguyên tắc FAIR và các khía cạnh đối với một tập hợp (siêu) dữ liệu....61

TÓM TẮT

Những tiến bộ về công nghệ đã làm cho nghiên cứu và khoa học ngày càng tăng cường dữ liệu hơn và kết nối lẫn nhau hơn, với các nhà nghiên cứu sản xuất và chia sẻ lượng dữ liệu ngày càng gia tăng. Trong nỗ lực của họ để sản xuất dữ liệu chất lượng cao, các nhà nghiên cứu phải tuân thủ các thực hành quản lý và quản trị dữ liệu tốt như các nguyên tắc Tìm thấy được, Truy cập được, Tương hợp được, Sử dụng lại được - FAIR (Findable, Accessible, Interoperable, Reusable).

Tuy nhiên, đã không có phân tích kỹ lưỡng để xác định giá trị của việc không có dữ liệu nghiên cứu FAIR, trong khắp các ngành khoa học, cả về các khía cạnh kinh tế và phi kinh tế, và để đối sánh nó với hiện trạng nơi đa số các dữ liệu nghiên cứu không gắn với các nguyên tắc FAIR. Báo cáo này nhằm lấp đi các khoảng trống đó bằng việc ước lượng chi phí của việc không có dữ liệu nghiên cứu FAIR đối với thị trường dữ liệu của Liên minh châu Âu và nền kinh tế dữ liệu của Liên minh châu Âu.

Phân tích của chúng tôi dựa vào các nghiên cứu có sẵn đã tập trung vào giá trị định lượng của dữ liệu nghiên cứu mà là tìm thấy được, truy cập được, tương hợp được, và sử dụng lại được.

Bằng việc xem xét tác động của FAIR lên các hoạt động nghiên cứu, cộng tác và đổi mới sáng tạo, các chỉ số đã được xác định, được định nghĩa và sau đó được định lượng. Bảy chỉ số đã được xác định để ước lượng chi phí của việc không có dữ liệu nghiên cứu FAIR: Thời gian bỏ ra, chi phí lưu trữ, chi phí cấp phép, rút lại nghiên cứu, cấp vốn hai lần, liên ngành và tăng trưởng kinh tế tiềm năng.

Để ước lượng được 5 chỉ số đầu tiên, chúng tôi trước hết đã đánh giá sự không hiệu quả nảy sinh trong các hoạt động nghiên cứu vì thiếu dữ liệu FAIR. Từ các mức không hiệu quả khác nhau, chúng tôi đã tính toán thời giản bị lãng phí vì không có FAIR và các chi phí có liên quan. Thứ hai, chúng tôi đã ước lượng chi phí các giấy phép bổ sung mà các nhà nghiên cứu phải trả tiền để truy cập dữ liệu mà lẽ ra sẽ là mở với các nguyên tắc FAIR. Thứ ba, chúng tôi đã xem xét các chi phí lưu trữ bổ sung vì thiếu dữ liệu FAIR. Dữ liệu không truy cập được dẫn tới việc tạo ra các bản sao dữ liệu bổ sung (ví dụ, bởi các tạp chí hoặc các trường đại học đối tác) điều lẽ ra là không cần thiết nếu các nguyên

tắc FARI được tuân thủ. Với dữ liệu không đủ để ước lượng 2 chỉ số cuối cùng, thay vào đó, chúng tôi cung cấp hầu hết các phát hiện và cân nhắc định tính.

Bám theo cách tiếp cận này, chúng tôi thấy chi phí thường niên của việc không có dữ liệu nghiên cứu FAIR khiến nền kinh tế châu Âu phải trả giá ít nhất 10.2 tỷ € mỗi năm. Ngoài ra, chúng tôi cũng đã liệt kê một số hệ lụy từ việc không có FAIR nhưng không thể ước tính được chính xác, ví dụ như tác động lên chất lượng nghiên cứu, doanh thu của nền kinh tế, hoặc khả năng máy đọc được các dữ liệu nghiên cứu. Bằng cách so sánh sơ bộ với nền kinh tế dữ liệu mở châu Âu, chúng tôi đã kết luận rằng các yếu tố không định lượng này có thể chiếm thêm 16 tỷ € hàng năm ngoài những gì chúng tôi đã ước tính. Các kết quả đó dựa vào sự kết hợp nghiên cứu bàn, các cuộc phỏng vấn với các chuyên gia theo từng vấn đề chủ đề và các giả định bảo thủ nhất của chúng tôi.

Ngoài ra, trong khi xây dựng dựa vào các nghiên cứu có sẵn và tin tưởng nhiều vào các tư liệu hiện có, chúng tôi đã tự nhận ra tầm quan trọng của việc có dữ liệu nghiên cứu FAIR. Không chỉ thời gian được đầu tư vào nghiên cứu này có thể được giảm thiểu với một lượng đáng kể, mà nội dung cũng có thể được cải thiện nếu nhiều tư liệu hơn là truy cập được và sử dụng lại được.

Cuối cùng, bằng việc ước lượng các chi phí định tính và định lượng của việc không có dữ liệu FAIR, báo cáo này sẽ xúc tác cho các nhà hoạch định chính sách đưa ra các quyết định dựa vào bằng chứng về các cách thức hiệu quả để hỗ trợ cho triển khai trong đời sống thực các nguyên tắc dữ liệu FAIR. Các nhà nghiên cứu và các cơ sở nghiên cứu bây giờ sẽ có khả năng cân nhắc chi phí của việc không có FAIR so với chi phí triển khai các nguyên tắc FAIR.

1. GIỚI THIỆU

1.1 Bối cảnh

Các tiến bộ công nghệ đã làm cho nghiên cứu và khoa học ngày càng tăng cường dữ liệu hơn và có kết nối lẫn nhau hơn, với các nhà nghiên cứu sản xuất và chia sẻ lượng dữ liệu ngày càng gia tăng. Trong nỗ lực của họ để sản xuất ra dữ liệu chất lượng cao, các nhà nghiên cứu phải tuân theo các thực hành quản lý dữ liệu và quản trị dữ liệu tốt. Ngoài việc thu thập, chú giải và lưu trữ đúng, quản lý và quản trị dữ liệu tốt bao gồm khái niệm chăm sóc lâu dài các tài sản kỹ thuật số có giá trị, dù đứng một mình hay trong sự kết hợp với các dữ liệu mới được sinh ra. Quản lý và quản trị dữ liệu tốt, bản thân nó, không phải là mục tiêu, mà thay vào đó, là con đường chính dẫn tới việc phát hiện và đánh giá dữ liệu và kiến thức dễ dàng hơn và đơn giản hơn, đồng thời dẫn tới việc tích hợp và sử dụng lại dữ liệu và kiến thức trong các nghiên cứu tiếp theo.

Để tối đa hóa giá trị của khoa học, dữ liệu nghiên cứu cần có 4 đặc tính cơ bản[1]; chúng cần phải là:

Tìm thấy được **F**indable	Truy cập được **A**ccessible	Tương hợp được **I**nteroperable	Sử dụng lại được **R**eusable
phát hiện được bằng siêu dữ liệu máy đọc được, nhận diện được và định vị được bằng phương tiện của một cơ chế nhận diện tiêu chuẩn	sẵn sàng và có thể có được cho cả con người và máy	vừa gói được về cú pháp, vừa hiểu được về ngữ nghĩa, cho phép trao đổi và sử dụng lại dữ liệu giữa các ngành khoa học, nhà nghiên cứu, cơ sở, tổ chức và quốc gia	được mô tả và chia sẻ hiệu quả bằng các giấy phép ít hạn chế nhất, cho phép sử dụng lại rộng rãi nhất có thể xuyên khắp các ngành khoa học và biên giới, và tích hợp ít trở ngại nhất với các nguồn dữ liệu khác

Hình 1: Bốn đặc tính cơ bản của FAIR

Tìm thấy được, Truy cập được, Tương hợp được và Sử dụng lại được - các nguyên tắc FAIR - có ý định để xác định một tập hợp tối thiểu các nguyên tắc và thực hành hướng

1 FAIR Principles described by GO-FAIR, https://www.go-fair.org/fair-principles/; FAIR Principles described by Force 11, https://www.force11.org/group/fairgroup/fairprinciples; and (Wilkison, Dumontier, & Mons, 2016).

dẫn có liên quan nhưng độc lập và tách biệt nhau, cho phép cả con người và máy tìm thấy, truy cập, tương hợp và sử dụng lại dữ liệu và siêu dữ liệu nghiên cứu.

1.1.1 Giá trị của FAIR

Triển khai các nguyên tắc FAIR có thể mang lại những lợi ích trực tiếp và gián tiếp cho các bên liên quan nghiên cứu, từ các nhà cấp vốn cho tới các nhà nghiên cứu, và có thể có tác động tích cực lên chất lượng và hoàn vốn đầu tư (ROI) của bản thân nghiên cứu. Các nghiên cứu hỗ trợ triển khai các nguyên tắc FAIR nói về tác động tích cực của chúng lên việc:

- Giảm đúp bản trong nghiên cứu, về các khía cạnh thời gian, nỗ lực và cấp vốn;
- Quản lý và quản trị nghiêm ngặt nguồn kỹ thuật số giúp cho các nhà nghiên cứu gắn với các kỳ vọng và yêu cầu của các cơ quan cấp vốn của họ[2];
- Mở rộng phạm vi các phát hiện nghiên cứu dựa vào dữ liệu sẵn có được tích hợp và được phân tích từ nhiều ngành và khu vực[3];
- Cho phép nghiên cứu tập trung nhiều hơn vào các hoạt động giá trị gia tăng như việc giải nghĩa dữ liệu thay vì tập trung vào việc tìm kiếm, thu thập hoặc tái tạo lại dữ liệu đang có[4]; và
- Cải thiện hạ tầng khoa học để hỗ trợ phát hiện kiến thức và đổi mới sáng tạo.

1.1.2 Các sáng kiến FAIR hiện hành

Nhiều sáng kiến hiện đang làm việc hướng tới triển khai các nguyên tắc FAIR. Chúng tôi liệt kê những sáng kiến chính bên dưới:

- Phong trào GO FAIR[5] (ĐI VỚI FAIR) tập hợp các sáng kiến quốc gia hiện có đã cam kết triển khai các nguyên tắc FAIR. Phong trào GO FAIR tiến hành các lựa chọn tập thể để triển khai về các khía cạnh tiêu chuẩn, thực hành tốt và các giao thức và đề xuất hạ tầng kỹ thuật, quản lý thay đổi và đào tạo quản trị dữ liệu.

2 (Wilkison, Dumontier, & Mons, 2016)
3 (Bonino da Silva Santos, et al., 2016)
4 http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf
5 https://www.go-fair.org/

- Thừa nhận tầm quan trọng của hệ sinh thái dữ liệu được FAIR xúc tác nhằm triển khai Đám mây Khoa học Mở châu Âu - EOSC (European Open Science Cloud)[6], Ủy ban châu Âu - cùng với các bên liên quan và được Nhóm Chuyên gia Dữ liệu FAIR[7] hỗ trợ - đang làm việc hướng tới một kế hoạch nhằm làm cho dữ liệu nghiên cứu thành FAIR và cho phép các nhà nghiên cứu có khả năng thực hiện việc chia sẻ và sử dụng lại dữ liệu khoa học liên ngành. Kế hoạch hành động này (còn được gọi là kế hoạch hành động dữ liệu FAIR) bao trùm tất cả các dạng đối tượng nghiên cứu kỹ thuật số và nó là một công cụ cộng tác hướng dẫn tích hợp các nguyên tắc FAIR ở mức châu Âu, xuyên biên giới và/hoặc các ngành. Ngoài ra, kế hoạch hành động dữ liệu FAIR tạo ra các điều kiện cho sự phát triển các kế hoạch đặc thù quốc gia và/hoặc ngành một cách mạch lạc để triển khai các nguyên tắc FAIR ở mức quốc gia và/hoặc ngành.
- Nghiên cứu về việc triển khai có khả năng các nguyên tắc FAIR ở Đan Mạch, nơi được Cơ quan về Khoa học và Giáo dục Đại học Đan Mạch ủy quyền. Nghiên cứu này chào một phân tích chi phí - lợi ích để triển khai trong tương lai các nguyên tắc FAIR cho nghiên cứu được nhà nước cấp vốn ở Đan Mạch.
- Nhiều trường đại học và viện nghiên cứu đã công khai ôm lấy phong trào FAIR và đang quảng bá cho các nguyên tắc FAIR. Các ví dụ bao gồm Dutch Techcentre for Life Science[8], Đại học Leiden[9], Đại học Utrecht[10], và Đại học Maastricht[11].
- ELIXIR là một hạ tầng phân tán về thông tin khoa học đời sống. ELIXIR[12] cam kết xúc tác cho tính sẵn sàng dữ liệu nghiên cứu FAIR trong khung EOSC[13]. ELIXIR Nodes cùng với EMBL-EBI[14], phối hợp và tích hợp các tài nguyên tin sinh học

6 European Cloud Initiative - Building a competitive data and knowledge economy in Europe (COM(2016) 178 final)
7 The Commission has established the Expert Group on FAIR data to support the Research and Innovation policy development on Open Science.
8 https://www.dtls.nl/fair-data/fair-data/
9 https://www.universiteitleiden.nl/en/research-dossiers/data-science/leiden-silicon-valley-of-fair-data
10 https://www.uu.nl/en/research/research-data-management/guides/costs-of-data-management
11 https://www.maastrichtuniversity.nl/research/data-science-um/research
12 ELIXIR Is a distributed infrastructure comprising 180 leading universities and centres of excellence
13 https://www.elixir-europe.org/system/files/elixir_statement_on_fair_data_management.pdf
14 European Molecular Biology Laboratory (EMBL) – European Bioinformatics Institute (EBI)

khắp các quốc gia thành viên (ví dụ, bằng việc cung cấp các cơ sở dữ liệu, các công cụ phân tích, các dịch vụ về tính tương hợp, .v.v.) với mục đích cuối cùng tạo ra thông tin sẵn sàng tự do không mất tiền cho cộng đồng khoa học.

1.1.3 Các thách thức triển khai các nguyên tắc FAIR

Thực tế là các nguyên tắc FAIR còn chưa là thực hành phổ biến vì vô số lý do. Vài nơi lo ngại việc thiếu nhận thức trong cộng đồng nghiên cứu[15] về cách để dữ liệu được chia sẻ, ở định dạng nào, thông tin hay siêu dữ liệu nào nên được cung cấp .v.v. Những nơi khác thấy động chạm tới các nền văn hóa hiện có và các hành vi trong tiến hành nghiên cứu, từ các nhà cấp vốn nghiên cứu cấm các nhà nghiên cứu chia sẻ dữ liệu của họ cho tới các nhà nghiên cứu thậm chí không coi dữ liệu họ sản xuất ra có thể là có giá trị cho những người khác, thiếu chú ý trong việc chuẩn bị kế hoạch quản lý dữ liệu, không có siêu dữ liệu[16], các tiêu chuẩn cạnh tranh khác nhau cho dữ liệu và siêu dữ liệu nghiên cứu, và thiếu các mã nhận diện thường trực cho dữ liệu, tập hợp dữ liệu và siêu dữ liệu[17].

Ngoài ra, nhiều nhà nghiên cứu và các tổ chức vẫn còn miễn cưỡng áp dụng các nguyên tắc FAIR và chia sẻ các tập hợp dữ liệu của họ vì các chi phí thực hoặc được thừa nhận, bao gồm việc đầu tư thời gian và tiền bạc.

Để dẫn dắt triển khai các nguyên tắc FAIR ở châu Âu, Ủy ban châu Âu cùng với một số các bên liên quan nghiên cứu tiên phong ở châu Âu đang tiến hành các hoạt động và và biện pháp nhằm nâng cao nhận thức về các chi phí và lợi ích của dữ liệu FAIR, và đang khuyến khích các cơ quan cấp vốn thiết lập các hướng dẫn hoặc hỗ trợ phát triển hạ tầng cho việc xuất bản dữ liệu FAIR.

Nghiên cứu hiện hành này đặt chính xác trong bối cảnh này. Dù thảo luận rất thường thấy xoay quanh các chi phí và lợi ích của việc triển khai các nguyên tắc FAIR cả về các khía cạnh kinh tế và phi kinh tế (và thậm chí có quan điểm toàn cầu vẫn còn đang được phát triển), ít ai nghĩ về các chi phí hiện hành và việc đánh mất các cơ hội từ việc không

15 Interview with Barend Mons, 2018-01-17

16 (Zahedi, Haustein, & Bowman, 2014), and (Parsons, Grimshaw, & Williamson, 2013)

17 (Johnson, Parsons, Chiarelli, & Kaye, JISC Research Data Assessment Support - Findings of the 2016 data assessment framework (DAF) surveys, 2016), Stehouwer & Wittenburg, 2014 and Tenopir C. , et al., 2011

triển khai các nguyên tắc FAIR (hoặc có rất ít các triển khai một phần nhỏ trong các ngành khoa học và các quốc gia). Đánh giá chi phí của việc không có dữ liệu FAIR sẽ cung cấp các số liệu định lượng để hỗ trợ cho việc sử dụng các nguyên tắc FAIR và sẽ giúp thuyết phục các bên liên quan nghiên cứu đầu tư vào triển khai của họ.

Nghiên cứu này trình bày và áp dụng phương pháp luận định lượng để ước lượng chi phí không có dữ liệu nghiên cứu FAIR ở châu Âu cũng như các ước lượng về chi phí. Phương pháp luận này được viết tốt thành tài liệu, dễ áp dụng ở các quốc gia và các ngành khác nhau và vì thế dễ được lặp lại trong tương lai.

Phạm vi của nghiên cứu này gồm các chi phí trực tiếp và gián tiếp là kết quả từ việc không triển khai các nguyên tắc FAIR ở châu Âu. Điều này bao gồm nghiên cứu được các tổ chức công, tư và phi chính phủ triển khai. Vì lý do này, báo cáo đã không xem xét chi phí triển khai dữ liệu nghiên cứu FAIR ở châu Âu cũng như không xem xét với việc mất mát doanh thu tiềm ẩn mà các nhà nghiên cứu phải gánh chịu như là kết quả của việc làm cho dữ liệu tự do không mất tiền và mở.

1.2 Khán thính phòng được mong đợi

Báo cáo này chủ yếu có ý định dành cho:

(1) Các nhà cấp vốn nghiên cứu: để nâng cao nhận thức về chi phí không có dữ liệu nghiên cứu FAIR, ví dụ, việc cấp vốn cho công việc dư thừa vì nghiên cứu trước đó không là FAIR. Với báo cáo này, các nhà cấp vốn có thể đánh giá tốt hơn tầm quan trọng của FAIR và đưa vào bất kỳ nơi nào thích hợp sự tuân thủ với các nguyên tắc FAIR như là điều kiện cho các nhà nghiên cứu để được vốn cấp.

(2) Các hạ tầng (dữ liệu) nghiên cứu: để nâng cao nhận thức về tác động của việc không có dữ liệu nghiên cứu FAIR, ví dụ, lên chi phí lưu trữ, và các hoạt động khác của các hạ tầng nghiên cứu.

(3) Các tổ chức thực hiện nghiên cứu: để nâng cao nhận thức về chi phí của việc không có dữ liệu nghiên cứu FAIR cho các nhà nghiên cứu, ví dụ, về các khía cạnh chất lượng nghiên cứu, thời gian bổ sung thêm vào việc nghiên cứu, và các trích dẫn. Bằng việc xác định chi phí không có dữ liệu nghiên cứu FAIR, báo cáo này giúp cho các nhà nghiên cứu để biện minh cho việc đầu tư vào việc tuân thủ với các nguyên tắc FAIR.

1.3 Dữ liệu nghiên cứu FAIR so với dữ liệu nghiên cứu mở

Nghiên cứu này áp dụng định nghĩa dữ liệu nghiên cứu của Chương trình H2020[18], ví dụ, thông tin, cụ thể là các sự kiện hoặc con số, được thu thập để được kiểm tra và xem xét như là cơ sở lý luận, thảo luận, hoặc tính toán. Các loại dữ liệu được đề cập trong định nghĩa này trước tiên là dữ liệu cơ bản (dữ liệu cần thiết để xác thực các kết quả được trình bày trong các xuất bản phẩm khoa học), bao gồm siêu dữ liệu liên quan (tức là siêu dữ liệu mô tả dữ liệu nghiên cứu được ký gửi) và thứ hai là bất kỳ dữ liệu nào khác (ví dụ: dữ liệu được giám tuyển không liên quan trực tiếp đến một xuất bản phẩm hoặc dữ liệu thô) cũng bao gồm cả siêu dữ liệu liên quan.

Dữ liệ FAIR và dữ liệu mở là 2 khái niệm riêng biệt, tuy nhiên, chúng ngày càng tiến gần nhau hơn. (Mons, et al., 2017) phân biệt khái niệm FAIR với khái niệm mở, nói: “Trong Internet của Dữ liệu và Dịch vụ FAIR được hình dung, mức độ ở đó bất kỳ mẩu dữ liệu nào cũng là sẵn có, hoặc thậm chí được quảng cáo như là đang sẵn có (thông qua siêu dữ liệu của nó) là hoàn toàn theo toàn quyền quyết định của chủ sở hữu dữ liệu đó”. Lý do đằng sau là “FAIR” chỉ nói về sự cần thiết phải:

- Mô tả một quy trình con người hoặc máy đọc được để truy cập dữ liệu được phát hiện;
- Mô tả rõ ràng và phong phú bối cảnh ở đó các dữ liệu đó đã được sinh ra, để xúc tác cho đánh giá tính khả dụng của nó;
- Linh hoạt xác định các điều kiện theo đó chúng có thể được sử dụng lại; và
- Cung cấp các chỉ dẫn rõ ràng về làm thế nào chúng sẽ được trích dẫn khi được sử dụng lại[19].

Tuy nhiên, nghiên cứu, dữ liệu sẽ không thực sự sử dụng lại được trừ phi nó là mở, nghĩa là, sẵn sàng theo một giấy phép mở và với các chi phí tượng trưng (trong hầu hết các trường hợp chi phí bằng 0), và tính mở thường đi cùng với sự triển khai các nguyên tắc FAIR. Như một phần của Chương trình nghị sự Khoa học Mở, Ủy ban châu Âu đã xác định tham vọng làm cho FAIR và truy cập mở thành mặc định cho việc chia sẻ dữ liệu

18 (Directorate-General for Research & Innovation (European Commission), 2017)

19 (Data Citation Synthesis Group, 2014)

trong nghiên cứu khoa học. Ngoài ra, Hội đồng của Liên minh châu Âu đã kết luận[20] rằng "càng mở càng tốt, đóng chỉ khi cần" là nguyên tắc cơ bản để sử dụng lại dữ liệu nghiên cứu một cách tối ưu.

Trong nghiên cứu này, chúng tôi vì thế coi chi phí của việc không có FAIR và truy cập miễn phí dữ liệu nghiên cứu bất cứ ở đâu thảo luận về các nguyên tắc FAIR.

1.4 Cấu trúc báo cáo

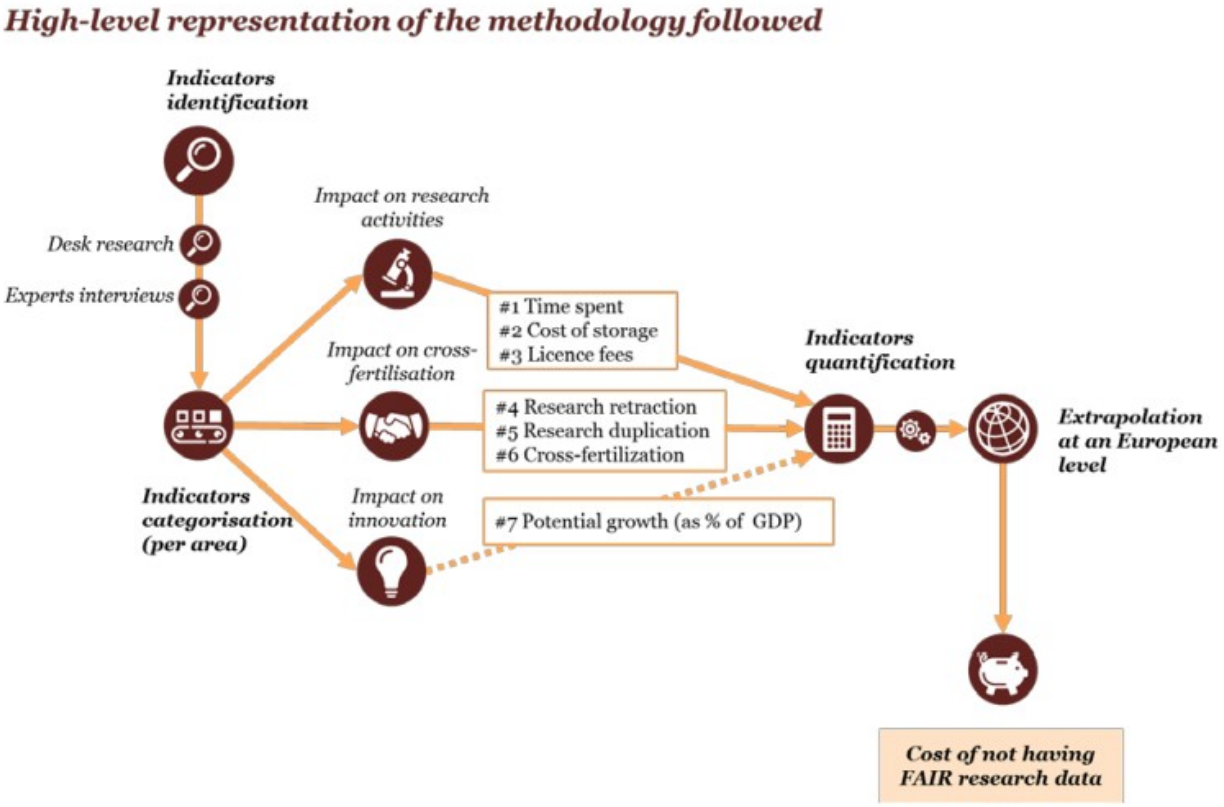
Phần còn lại của nghiên cứu này có cấu trúc như sau:

- Chương 2 "Cách tiếp cận" đưa ra cách tiếp cận của chúng tôi nhằm xác định các chỉ số đo lường chi phí của việc không có dữ liệu nghiên cứu, mô tả từng chỉ số và phương pháp luận định lượng chúng.
- Chương 3 "Tính toán chi phí" trình bày các kết quả bài tập định lượng chi phí không có dữ liệu FAIR.
- Chương 4 "Kết luận" thảo luận các phát hiện và quan sát chính của nghiên cứu.

20 (Council of the European Union, 2016)

2. CÁCH TIẾP CẬN

2.1 Tổng quan



Hình 2: Trình bày mức cao phương pháp luận

Trong nghiên cứu này, chúng tôi đã bắt đầu bằng việc xác định và lựa chọn các chỉ số kinh tế dẫn tới chi phí của việc không có dữ liệu nghiên cứu FAIR. Các chỉ số đó - được kết nối với các mệnh đề về chi phí xảy ra vì không có dữ liệu nghiên cứu FAIR - đã được phân loại và định lượng tới mức có thể. Ngoài ra, cách tiếp cận này chỉ dựa vào dữ liệu thứ cấp[21].

Để định lượng các chỉ số đó, các giả thiết và khẳng định đã được đưa ra để cho phép chúng tôi ước lượng chi phí của việc không có dữ liệu nghiên cứu FAIR. Phương pháp ước lượng các chỉ số và các giả thiết cần thiết được mô tả.

Các tiêu chí khác nhau đã được xem xét nhằm xác định và định lượng các chỉ số:

21 Secondary data refers to information that already exist, in our case information that we collected rather than created.

- Độ chính xác: chỉ số này có liên quan tới các hoạt động cụ thể được xác định theo các điều khoản rõ ràng;
- Độ tin cậy: chỉ số này dựa vào các sự việc và đo lường được nhất quán theo thời gian, điều đảm bảo tính bền vững của ước lượng này;
- Khả năng đo lường: chỉ số này có thể được định lượng bằng việc sử dụng các công cụ và phương pháp hiện có; và
- Không chồng lấn: từng chỉ số có thể được đo lường độc lập với các chỉ số khác.

2.2 Xác định các chỉ số

Để đo lường chi phí của việc không tuân thủ các nguyên tắc FAIR, chúng tôi trước hết cần đánh giá tác động của chúng lên dữ liệu nghiên cứu. Vì thế, chúng tôi đã chi tiết hóa trong Phụ lục I các khía cạnh khác nhau của từng nguyên tắc, và đã rà soát lại các thước đo được các tác phẩm có liên quan gần đây đề xuất[22] cho việc ước lượng sự tuân thủ dữ liệu nghiên cứu theo các nguyên tắc FAIR. Các khía cạnh và các thước đo đó được sử dụng để xác định các chỉ số kinh tế được kết nối tới các dữ liệu nghiên cứu không FAIR. Các chỉ số kinh tế không liên quan tới các nguyên tắc FAIR đã bị loại bỏ.

Chúng tôi dự kiến tác động của các nguyên tắc FARI lên dữ liệu nghiên cứu đối với các nhà nghiên cứu, các kho và các nhà xuất bản, và đổi mới sáng tạo. Việc xác định các chỉ số kinh tế về các chi phí không có dữ liệu nghiên cứu FAIR cũng hưởng lợi từ nghiên cứu bàn và việc phỏng vấn các chuyên gia.

Ba lĩnh vực sau đây nhóm lại các chỉ số kinh tế khác nhau được xác định như một phần của chi phí không có dữ liệu nghiên cứu FAIR. Các lĩnh vực đó là áp dụng được cho tất cả các khu vực (ví dụ, hàn lâm, tư nhân, nhà nước, phi lợi nhuận).

- Tác động lên các hoạt động nghiên cứu: Chủng loại này của các chỉ số liên quan tới các hoạt động diễn ra trong quá trình nghiên cứu. Nó liên quan, nhưng không bị hạn chế tới, phát triển nghiên cứu, lập kế hoạch dự án nghiên cứu, tạo lập dữ

22 (Dunning, de Smaele, & Böhmer, 2017)
https://via.hypothes.is/https://www.force11.org/group/fairgroup/fairprinciples
http://datafairport.org/fair-principles-living-document-menu (Wilkison, Dumontier, & Mons, 2016)
https://www.force11.org/group/fairgroup/fairprinciples
https://via.hypothes.is/https:/content.iospress.com/articles/information-services-and-use/su824#ref019

liệu, thu thập dữ liệu, tiền xử lý và làm sạch dữ liệu, tích hợp và chuyển đổi dữ liệu, .v.v. Trong các hoạt động đó, thời gian bị lãng phí vì dữ liệu nghiên cứu không là FAIR được xác định và định lượng. Chủng loại này cũng bao gồm sự đúp bản lưu trữ dữ liệu nghiên cứu không là FAIR (ví dụ, dữ liệu nghiên cứu đang được lưu trữ không cần thiết trong các kho khác nhau). Sự đúp bản này được định lượng về các khía cạnh chi phí lưu trữ. Ngoài ra, chúng tôi cũng đưa vào chủng loại này các chi phí các nhà nghiên cứu đối mặt trong việc tìm kiếm, truy cập và phân tích dữ liệu nghiên cứu mà không là mở. Ví dụ, thời gian bỏ ra để có được dữ liệu (như, quy trình đăng ký), các khoản phí và các chi phí cấp phép có liên quan, .v.v.

- Tác động lên sự cộng tác: đây là chi phí cơ hội đối với cộng đồng nghiên cứu nhưng cũng cả cho các cá nhân và các tổ chức sử dụng dữ liệu nghiên cứu. Chi phí này thể hiện dưới dạng bỏ lỡ các cơ hội hợp tác hoặc sinh lợi chéo giữa các nhóm nghiên cứu trong và giữa các ngành. Là quan trọng để lưu ý rằng ước lượng chi phí này rất nhạy cảm đối với những giả thiết được đưa ra.
- Tác động lên đổi mới sáng tạo: Chủng loại này liên quan tới tác động của dữ liệu nghiên cứu không là FAIR lên đổi mới sáng tạo và hệ quả là lên nền kinh tế của châu Âu. Nhiều bằng chứng thực nghiệm chứng minh rằng nghiên cứu là “động lực chính của năng suất và tăng trưởng kinh tế”[23]. Tương tự, dữ liệu nghiên cứu không là FAIR có thể có tác động lên số lượng các bằng sáng chế được đệ trình, các cơ hội bị bỏ lỡ về các khía cạnh kinh doanh hoặc các sản phẩm mới, tạo ra công việc ít hơn, .v.v.

Trong phần còn lại của nghiên cứu này, chúng tôi phân biệt giữa các nhà nghiên cứu hàn lâm và phi hàn lâm, cái sau xoay quanh các khu vực tư, công và phi lợi nhuận. Đối với các nhân viên hàn lâm, khái niệm ‘nhà nghiên cứu’ bao gồm tất cả những người mà nghiên cứu đối với họ là một trong các hoạt động cốt lõi như các giáo sư, các nghiên cứu sinh tiến sỹ, sau tiến sỹ, trợ lý giáo sư, .v.v. Đối với các nhân viên phi hàn lâm, khái niệm ‘nhà nghiên cứu’ bao gồm tất cả những ai không phải là nhân viên hàn lâm nhưng đủ điều kiện là 'nghiên cứu chuyên sâu' (tức là hoạt động cốt lõi của họ là nghiên cứu).

23 (Directorate-General for Research and Innovation (European Commission), 2017)

2.2.1 Tác động lên các hoạt động nghiên cứu

Khi phân tích các nghiên cứu hiện hành về chi phí có liên quan tới các hoạt động nghiên cứu, chúng tôi đã xác định một tập hợp các hoạt động theo đó các chỉ số chi phí có thể được liên kết. Các hoạt động đó đã được biên soạn từ các tư liệu hiện hành về vòng đời dữ liệu[24] và các hoạt động nghiên cứu[25].

Hầu hết các chi phí các bên liên quan tới nghiên cứu đối mặt có thể được liên kết trực tiếp tới một hoặc nhiều hoạt động sau đây (cũng xem Hình 2):

- Phát triển nghiên cứu
- Lập kế hoạch một dự án nghiên cứu
- Tạo lập và thu thập dữ liệu
- Tiền xử lý và làm sạch dữ liệu
- Tích hợp dữ liệu
- Phân tích dữ liệu
- Biên tập báo cáo
- Đăng ký và xuất bản
- Rà soát lại ngang hàng

Các nguyên tắc FAIR không có tác động trực tiếp lên tất cả các hoạt động đó, điều này giải thích vì sao phát triển nghiên cứu, lập kế hoạch dự án nghiên cứu và biên tập báo cáo không nằm trong phân tích của chúng tôi.

24 (Data Life Cycle | DataONE, n.d.)
25 (Ziker, 2013)

Hình 3: Các hoạt động nghiên cứu

2.2.1.1 Tạo lập và thu thập dữ liệu

Tạo lập và thu thập dữ liệu là một bước trong vòng đời nghiên cứu nơi dữ liệu được tạo ra (nghĩa là thông qua các quan sát, thí nghiệm hoặc mô phỏng), và dữ liệu tiềm tàng hữu ích và hiện có tìm thấy được và giành được. Bước này cũng bao gồm việc thực hiện kiểm tra tính xác thực của dữ liệu.

Trong hoạt động này của vòng đời nghiên cứu, việc không có FAIR có tác động lên thời gian, các chi phí lưu trữ và các chi phí cấp phép cần thiết để tìm kiếm, truy cập, điều khiển và sử dụng lại (siêu) dữ liệu chống trụ cho các tài liệu nghiên cứu và các nghiên cứu còn chưa được làm cho sẵn sàng (đúng cách), vì các lý do khác nhau:

- Lãng phí thời gian vì không tìm thấy được, con người không hiểu được hoặc siêu dữ liệu không có cấu trúc hoặc không hoàn chỉnh. Nhiều nhà nghiên cứu không xuất bản một tập hợp tối thiểu siêu dữ liệu cho các nghiên cứu của họ, điều cần thiết cho các nhà nghiên cứu khác để tìm kiếm, truy cập và sử dụng lại nó. Siêu dữ liệu chất lượng tồi cản trở khả năng tìm thấy được. Ngoài ra, như được mô tả

trong Phụ lục I, siêu dữ liệu cần đủ phong phú cho các độc giả bên ngoài để trực tiếp hiểu thông tin cần thiết về nghiên cứu và dữ liệu có tính trợ giúp của nó.

- Chi phí về thời gian từ sự tinh thông của con người cần thiết để đọc và hiểu dữ liệu. Việc có dữ liệu được làm thành tài liệu tốt có thể dẫn tới tổng chi phí thấp hơn nhiều về các khía cạnh nắm bắt thông tin cơ bản bên trong tập hợp dữ liệu đó và làm việc với nó (ví dụ, hiểu đầy đủ về lĩnh vực nghiên cứu).
- Thiếu điểm truy cập duy nhất[26] đang làm gia tăng thời gian cần thiết đối với các nhà nghiên cứu để xác thực các kênh xuất bản khác nhau, phải tìm kiếm trong từng kênh, .v.v. Phong trào FAIR đang bênh vực cho điểm truy cập duy nhất, nơi bạn có thể truy vấn về siêu dữ liệu có thể nhất quán trỏ tới một tầng nơi dữ liệu đó nằm, vì thế giảm thiểu đáng kể thời gian truy cập (siêu) dữ liệu. Trong trường hợp dữ liệu hiện hành chưa sẵn sàng hỗ trợ cho các phát hiện đã được xuất bản, cần thiết có thời gian để tạo lại tập hợp dữ liệu.
- Dữ liệu có liên quan tới nghiên cứu nhất định hoặc toàn bộ ngành nghiên cứu thường được lưu trữ biệt lập, trong kho chứa thông tin khép kín. Thường thấy là dữ liệu một phần sẵn sàng từ các điểm truy cập khác nhau[27], dẫn tới việc nhà nghiên cứu phải bỏ ra nhiều thời gian hơn để truy xuất tất cả dữ liệu họ cần. Nhà nghiên cứu này cũng có thể cần chuyển dữ liệu từ kho này sang kho khác, điều có thể làm gia tăng tổng chi phí lưu trữ.
- Dữ liệu bị đúp bản khắp một hoặc nhiều kho làm cho các nhà nghiên cứu phải bỏ ra nhiều thời gian hơn để lựa chọn các tập hợp dữ liệu đúng.
- Việc làm cho dữ liệu nghiên cứu tuân thủ các nguyên tắc FAIR sẽ có tác động tích cực lên truy cập mở tới các xuất bản phẩm khoa học và hệ quả là sẽ làm giảm các khoản phí truy cập (nghĩa là, các chi phí cấp phép) các nhà nghiên cứu đối mặt, điều sẽ cải thiện tỷ lệ trích dẫn (nghĩa là, sử dụng lại) các xuất bản phẩm của họ[28].

26 E.g. EOSC ambitions to be “a one-stop-shop to find, access, and use research data and services from multiple disciplines and platforms.”

27 A single point of access should not be confused with a single repository. The first would only harvest and store the metadata while the second refers to all types of data.

28 (Van Noorden, 2013)

- Cuối cùng, siêu dữ liệu không có cấu trúc làm cho máy khó đọc được, đây là một trong những quan điểm được các nguyên tắc FAIR bênh vực. Siêu dữ liệu máy không đọc được và không tương hợp được làm cho các máy tìm kiếm không thể đánh chỉ mục được sẽ cản trở lớn khả năng tìm thấy được dữ liệu có liên quan[29].

Việc phân tích dữ liệu cho nguyên mẫu của máy thay vì vốn con người có thể làm giảm thời gian các nhà nghiên cứu bỏ ra và toàn bộ các chi phí có liên quan (ví dụ, các chi phí cấp phép để truy cập phần mềm nhất định, sự tinh thông từ bên ngoài, chuyển dữ liệu vì truy cập từ xa là không thể) tới nó xuống hầu như bằng không (zero). Tình trạng hiện hành ép các nhà nghiên cứu phải đọc và hiểu siêu dữ liệu một cách thủ công, và cản trở việc tìm kiếm và phát hiện dữ liệu được tự động hóa.

Ngoài ra, các khó khăn trong việc đánh giá chất lượng và tính toàn vẹn dữ liệu vì (siêu) dữ liệu không đầy đủ làm cho các nhà nghiên cứu không thể chia sẻ và phân tích dữ liệu nghiên cứu trong một môi trường tin cậy xuyên khắp các công nghệ, liên ngành và xuyên biên giới.

2.2.1.2 Tiền xử lý và làm sạch dữ liệu

Tiền xử lý và mô tả dữ liệu là một bước trong vòng đời nghiên cứu nơi việc thẩm định và kiểm tra được thực hiện để cải thiện và đảm bảo chất lượng dữ liệu. Việc mô hình hóa dữ liệu và làm sạch dữ liệu cũng diễn ra trong bước này.

Việc không có FAIR có tác động lên thời gian cần thiết để tạo lập, giám tuyển, làm sạch, xử lý lại và giữ cho dữ liệu và siêu dữ liệu được cập nhật. Việc duy trì siêu dữ liệu chất lượng thấp phát sinh chi phí đáng kể và có thể là nguồn gây ra lỗi. Chi phí này sẽ gia tăng với thời gian vì nhiều lý do tiềm tàng: các yếu tố con người, các tiêu chuẩn bị/được thay đổi, .v.v.

Ví dụ, thời gian là cần thiết để tiến hành kiểm tra chất lượng (ví dụ, chạy các script/các truy vấn) trên cả các dữ liệu thô và dữ liệu hiện đang được thu thập. Tương tự, thời gian là cần thiết để cải thiện chất lượng (siêu) dữ liệu và để chuyển đổi dữ liệu khi cần thiết (ví dụ, để xác định và sửa các lỗi).

Dữ liệu nghiên cứu FAIR sẽ là máy đọc được, vì thế làm giảm thời gian cần thiết cho việc kiểm tra chất lượng. Các lý do gồm:

29 (Stehouwer & Wittenburg, 2014)

- Các phương pháp luận giám tuyển thường được phát minh vào thời điểm cần thiết và hệ quả là không thể nhân bản được cho siêu dữ liệu khác, điều dẫn tới không hiệu quả trong tương lai, và
- Thiếu tài liệu về tính không đồng nhất rộng rãi của các hoạt động tiền xử lý đang được triển khai, tùy thuộc vào dữ liệu đã được tạo ra.

2.2.1.3 Tích hợp dữ liệu

Tích hợp dữ liệu là hoạt động trong vòng đời nghiên cứu nơi dữ liệu từ các nguồn khác nhau được tổng hợp để tạo thành một tập hợp dữ liệu đồng nhất có thể phân tích được. Vì lượng dữ liệu nghiên cứu đang gia tăng[30], tích hợp dữ liệu trở nên cần thiết để có khả năng tổng hợp dữ liệu từ nhiều nguồn (và từ các định dạng khác nhau).

Việc không có FAIR có tác động lên thời gian cần thiết để đối sánh/chuyển đổi hai hoặc nhiều hơn các tập hợp dữ liệu trong một định dạng được hài hòa hóa và máy đọc được (từ các định dạng sở hữu độc quyền sang các định dạng được tiêu chuẩn hóa), hoặc thời gian cần thiết để hiểu rằng hai tập hợp dữ liệu khác nhau không thể tích hợp được vì các từ vựng không tương hợp (ví dụ, siêu dữ liệu) được sử dụng để xác định các tập hợp dữ liệu. Như một quy tắc, trong một dự án phân tích dữ liệu, việc làm sạch dữ liệu đối với các dữ liệu chất lượng tồi có thể chiếm tới 80% tổng nỗ lực.

Cần lưu ý là việc lấp đi khoảng trống giữa các phân loại ngữ nghĩa và các chú giải là một thách thức vì bản chất tự nhiên năng động của nghiên cứu, sự năng động của các phân loại và các công cụ khó dùng[31].

Trong khu vực tư nhân, các chi phí tích hợp đặc biệt cao trong trường hợp liên danh nghiên cứu chung, vì thời gian để tích hợp dữ liệu không đồng nhất là đáng kể[32]. Điều này đôi khi dẫn tới việc không sử dụng được các tập hợp dữ liệu.

Ngoài ra, các chuyên gia trong một vài cộng đồng đã phát triển các hướng dẫn về siêu dữ liệu của riêng họ. Trong khi các hướng dẫn cung cấp mức độ linh hoạt cao, và vì thế năng lực đại diện cũng cao, nó trước hết đòi hỏi các chuyên gia tận dụng chúng[33] và

30 Interview with Barend Mons, 2018-01-17
31 Peter Wittenburg
32 Interview with Barend Mons, 2018-01-17
33 Second year report on RDA Europe analysis programme

thứ hai, các hướng dẫn đó thường không phù hợp với các tiêu chuẩn được thừa nhận, gây khó cho khả năng đọc được của cả máy và con người.

2.2.1.4 Phân tích dữ liệu

Phân tích dữ liệu là hoạt động trong vòng đời nghiên cứu nơi dữ liệu được xử lý với mục đích cuối cùng là để truy xuất thông tin hữu ích, để tạo thuận lợi cho những thấu hiểu và cuối cùng để hình thành các quan sát và kết luận. Phân tích dữ liệu hoặc việc mô hình hóa dữ liệu thường được thực hiện bằng việc sử dụng các phần mềm, công cụ hoặc phương pháp luận nhất định.

Trong bước này của vòng đời nghiên cứu, việc không có FAIR có tác động lên thời gian cần thiết để phân tích dữ liệu không được ghi thành tài liệu tốt hoặc không hoàn chỉnh và ở mức chất lượng đúng. Theo một tài liệu gần đây, “có lo ngại ngày một gia tăng về khả năng nhân bản và tái tạo lại các kết quả nghiên cứu: một khảo sát gần đây đã chỉ ra rằng hơn 70% các nhà nghiên cứu đã cố gắng và thất bại trong việc tái tạo lại các kết quả của các nhà khoa học khác, và hơn một nửa đồng ý rằng có cuộc khủng hoảng đáng kể về khả năng tái tạo lại” (Baker, 2016). Điều đó cũng liên quan đến thời gian cần thiết để xác minh các phát hiện của các nghiên cứu khác nhằm quyết định liệu phương pháp luận và các kết quả có chính xác hay không và có thể là đối tượng của một xuất bản phẩm hay không.

2.2.1.5 Đăng ký và xuất bản

Đăng ký và xuất bản là một bước trong vòng đời nghiên cứu nơi dữ liệu được mô tả chính xác và đầy đủ bằng việc sử dụng siêu dữ liệu có liên quan. (Siêu) dữ liệu sau đó được đăng ký (ví dụ, được lưu trữ) trên trực tuyến, ví dụ, trên các máy chủ web và được xuất bản bằng việc sử dụng các kênh tương xứng. Việc không có FAIR có tác động lên thời gian và các chi phí lưu trữ, gồm:

- Thời gian đăng ký (siêu) dữ liệu và duy trì một kho nghiên cứu với (siêu) dữ liệu không là FAIR, hoặc để giữ cho một cơ sở dữ liệu nhất định được cập nhật với mức tương hợp nhất định.
- Các chi phí lưu trữ có liên quan tới lượng dữ liệu bị đúp bản trong các kho khác nhau, độc lập với cơ chế bảo tồn được sử dụng. Ví dụ, dữ liệu không truy cập được hoặc được mô tả sai có thể dẫn tới các đúp bản trong và giữa các kho.

Các quy trình và các yêu cầu xuất bản và đánh chỉ mục biến động khác nhau giữa các kho, làm gia tăng thời gian phải bỏ ra cho các hoạt động đó.

Khi không có dữ liệu và các xuất bản phẩm truy cập được, các tạp chí, các trường đại học và các nhà cấp vốn thường phải có một bản sao riêng của dữ liệu nghiên cứu trong một kho. Sự dư thừa không cần thiết này dẫn tới tổng các chi phí lưu trữ gia tăng.

2.2.1.6 Rà soát lại ngang hàng

Rà soát lại ngang hàng là một bước trong vòng đời nghiên cứu nơi công việc nghiên cứu được các nhà nghiên cứu khác đánh giá với sự tinh thông hoặc kinh nghiệm thích hợp trước khi công việc đó có thể được phê chuẩn, xuất bản hoặc trình bày.

Trong bước này của vòng đời nghiên cứu, việc không có FAIR có tác động lên thời gian bỏ ra cho các rà soát lại dư thừa. Các rà soát lại dữ liệu và xuất bản phẩm nghiên cứu hiếm khi được chia sẻ, và thậm chí khi chúng được chia sẻ, việc thiếu các tiêu chuẩn rộng khắp nền công nghiệp ngụ ý là mỗi thực thể trưng cầu sự rà soát lại riêng của mình trước khi đưa ra quyết định[34].

Ngoài ra, từng rà soát lại ngang hàng có thể là mất thời gian vì thiếu dữ liệu hỗ trợ cho các kết luận của bài báo được rà soát lại.

Theo hiệp hội thương mại dành cho các nhà xuất bản học thuật và chuyên nghiệp[35], rà soát lại ngang hàng chiếm 15 triệu giờ mỗi năm. Quy trình này trung bình cần tới 2 hoặc 3 người và tỷ lệ chấp nhận trung bình là khoảng 50%. Rà soát lại ngang hàng thường được coi là một công cụ đảm bảo chất lượng nhằm cải thiện chất lượng các công trình nghiên cứu. Tuy nhiên, quy trình này thường là dư thừa xuyên khắp cộng đồng khoa học và vì thế đôi khi là không cần thiết.

2.2.2 Tác động lên sự cộng tác

Cộng tác được áp dụng cho khoa học tất cả là về việc kết hợp dữ liệu và các phát hiện nghiên cứu từ các tổ chức khác nhau trong và xuyên các ngành để sản xuất ra các kết quả đầu ra mới và tốt hơn. Vì thế, có nhiều tác động vì dữ liệu nghiên cứu không là FAIR. Một khảo sát[36] được tiến hành ở Đại học Sheffield đã chỉ ra rằng 31% các nhân

34 (American Journal Experts, n.d.)
35 (Ware & Mabe, 2012)
36 (Cox & Williamson, 2014)

viên hàn lâm đồng ý với tuyên bố sau: “Thiếu truy cập tới dữ liệu được các nhà nghiên cứu hoặc các cơ sở khác tạo ra đã hạn chế khả năng của tôi để trả lời các câu hỏi nghiên cứu”. Một nghiên cứu khác[37] tham chiếu tới 50,1% những người trả lời nêu đã hạn chế khả năng trả lời các câu hỏi khoa học vì thiếu truy cập tới dữ liệu được các nhà nghiên cứu khác tạo ra.

Ngoài ra, một nghiên cứu gần đây[38] cho thấy 73% nhân viên hàn lâm được khảo sát nói rằng việc có truy cập tới dữ liệu nghiên cứu được xuất bản có thể làm lợi cho nghiên cứu của riêng họ. Tương quan với số liệu đó, những người trả lời khảo sát đã chỉ ra rằng việc chia sẻ dữ liệu nghiên cứu là quan trọng để tiến hành nghiên cứu trong lĩnh vực của họ. Việc có dữ liệu nghiên cứu FAIR vì thế có thể xúc tác cho sử dụng lại dữ liệu nghiên cứu và vì thế cho cộng tác. Sự cộng tác tốt hơn có thể tác động trực tiếp và tích cực tới lợi ích đan xen[39] và vì thế đổi mới sáng tạo.

Khả năng tìm thấy các báo cáo có khả năng tùy thuộc vào tính sẵn sàng của chúng trong các biên mục như Google Scholar và Scopus hoặc đặc biệt các tạp chí và các cổng nghiên cứu. Bằng việc gia tăng sử dụng các kênh phổ biến sẵn sàng công khai, FAIR làm gia tăng khả năng các báo cáo được tìm thấy và được sử dụng lại.

Ngoài ra, nhiều tài liệu vẫn không được trích dẫn[40] vì sự cộng tác yếu kém. Sử dụng dữ liệu cũng có thể có tác động khổng lồ lên trích dẫn và vì thế sự cộng tác và lợi ích đan xen như được một nghiên cứu gần đây chỉ ra[41]. Đã được chứng minh rằng các nghiên cứu mà dữ liệu được làm cho sẵn sàng trong một kho công khai nhận được nhiều trích dẫn hơn so với các nghiên cứu tương tự với dữ liệu không được làm cho sẵn sàng.

Việc triển khai các nguyên tắc FAIR có thể làm gia tăng một phần sự cộng tác giữa các cộng đồng khoa học. Ví dụ, một khảo sát được tiến hành ở Đại học Sheffield cho thấy 64% các nhà nghiên cứu được khảo sát có thiện chí chia sẻ dữ liệu với những người khác không có hạn chế thông qua một kho dữ liệu trung tâm. Tuy nhiên, việc định lượng tác động trực tiếp của FAIR lên cộng tác là thách thức vì có thể phải đánh giá giá

37 (Tenopir, et al., 2011)
38 https://www.elsevier.com/about/open-science/research-data/open-data-report
39 Cross-fertilization refers to the mixing of data from different disciplines to produce a better result.
40 (Davis, 2012)
41 (Piwowar & Vision, 2013)

trị của một nghiên cứu đã được làm với FAIR so với giá trị của nghiên cứu đã không được làm với FAIR.

Vì điều này, chúng tôi xem xét tác động của việc không có FAIR lên cộng tác bằng việc sử dụng các ủy quyền sau:

- Rút lại nghiên cứu: nghiên cứu bị rút lại như là hệ quả của việc không có FAIR không thể đóng góp cho sự tiến bộ của khoa học. Điều này bao gồm nghiên cứu bị rút lại vì các lỗi, không có khả năng tái tạo lại, giả mạo, đạo văn, và các lý do khác[42]. Chúng tôi mặc định rằng các nguyên tắc FAIR có thể làm giảm giả mạo và làm tăng chất lượng có thể quan sát thấy trong việc giảm số lượng các bài báo bị rút lại.
- Đúp bản (cấp vốn) nghiên cứu: nghiên cứu thừa không đóng góp cho khoa học.
- Lợi ích đan xen: nghiên cứu được làm cho có thể nhờ các nguyên tắc FAIR, điều có thể nếu khác sẽ là không thể (ví dụ, vì dữ liệu nó dựa vào có thể đã không sử dụng lại được).

2.2.3 Tác động lên đổi mới sáng tạo

Tác động lên đổi mới sáng tạo đại diện cho chi phí cơ hội của việc không có dữ liệu nghiên cứu FAIR. Trong một khảo sát được xuất bản về "Các nhà khoa học chia sẻ dữ liệu: Các thực hành và nhận thức"[43], 64% những người được hỏi đã trả lời rằng thiếu truy cập tới dữ liệu được các nhà nghiên cứu hoặc các cơ sở khác tạo ra là một cản trở chính đối với tiến bộ khoa học. Một nghiên cứu khác[44] ủng hộ ý tưởng này với số liệu hơn 43%, với khác biệt duy nhất là họ đã nêu cản trở chính cho sự tiến bộ trong ngành của họ. Từ các quan sát đó, chúng tôi kết luận rằng việc có dữ liệu nghiên cứu FAIR có thể có tác động tích cực lên đổi mới sáng tạo vì ngược lại được nêu làm chậm sự tiến bộ. Ngoài nghiên cứu và khoa học, có thể lý giải rằng các doanh nghiệp cũng đang phải đối mặt với những cạm bẫy vì tính sẵn sàng, việc sử dụng và kết hợp dữ liệu là rất quan trọng để phát triển các dịch vụ và sản phẩm mới và phân biệt với các dịch vụ và sản phẩm hiện có[45].

42 (Fang, Steen, & Casadevall, 2012) and (Wager & Williams, 2011)
43 (Tenopir, et al., 2011)
44 (Cox & Williamson, 2014)
45 (Wittenburg, Costs of FAIR Compliance and not being FAIR compliant, 2017)

Việc không triển khai các nguyên tắc FAIR tác động lên đổi mới sáng tạo theo các cách thức khác nhau, đây là một danh sách chưa vét cạn:

- Thiếu truy cập tới dữ liệu giá trị cao cản trở sự phát triển các dịch vụ đổi mới sáng tạo và tạo ra các mô hình kinh doanh mới.
- Dữ liệu nghiên cứu không FAIR cũng sẽ cản trở việc sử dụng Khoa học Máy (Machine Science)[46] và những gì nó kéo theo. Dữ liệu máy không đọc được, sẽ không có khả năng để máy xử lý lượng dữ liệu ngày một gia tăng và trích xuất các thấu hiểu ở cấp độ mà con người không thể làm được.
- Đổi mới sáng tạo và tiến bộ được làm cho có thể bằng việc xây dựng dựa vào các kết quả của công việc trước đó. Nếu dữ liệu nghiên cứu không sẵn sàng, nghiên cứu mới sẽ luôn bổ sung cho đường cơ sở y hệt, vì thế cản trở đổi mới sáng tạo.
- Không có khả năng dễ dàng xác định những người cộng tác, các đối tác và các chuyên gia nghiên cứu có thể vì dữ liệu nghiên cứu không là FAIR.
- Mất dữ liệu bắt nguồn từ quản lý dữ liệu yếu kém hoặc thiếu chính sách giữ lại rõ ràng.
- Thiếu rõ ràng về các giấy phép và các điều kiện sử dụng dữ liệu cản trở việc sử dụng dữ liệu không FAIR trong các dự án giá trị gia tăng, vì kiện tụng và các rủi ro vi phạm giấy phép.

Nhìn chung, tác động của FAIR lên đổi mới sáng tạo thể hiện ở các khía cạnh tăng trưởng kinh tế và tạo ra công ăn việc làm không hiện thực hóa được.

46 Machine Science is defined by the increasing use of computational resources in research-related activities. In our case, Machine Science specifically refers to machine-readability and reusability of data.

2.3 Các chỉ số được lựa chọn

Dựa vào 3 lĩnh vực ở trên và phân tích của chúng tôi, các chỉ số chúng tôi đã xác định là như sau:

Các lĩnh vực	Các chỉ số
Tác động lên các hoạt động nghiên cứu	1. Thời gian bỏ ra 2. Chi phí lưu trữ 3. Chi phí cấp phép
Tác động lên các cơ hội nghiên cứu tiếp	4. Rút lại nghiên cứu 5. Cấp vốn hai lần 6. Lợi ích đan xen (Liên ngành)
Tác động lên đổi mới sáng tạo	7. Tăng trưởng kinh tế tiềm năng (% GDP)

Bảng 1: Các chỉ số được lựa chọn cho từng lĩnh vực

Chỉ số #1: Thời gian bỏ ra

Để tính toán thời gian lãng phí trong quá trình các hoạt động nghiên cứu, chúng tôi ước lượng tổng thời gian các nhân viên nghiên cứu bỏ ra để tiến hành các hoạt động có liên quan tới nghiên cứu được liệt kê ở trên. Từ tổng thời gian đó chúng tôi đã dẫn xuất thời gian bị lãng phí vì dữ liệu không là FAIR và đã chuyển đổi nó thành giá trị tài chính bằng việc sử dụng lương trung bình cho các nhà nghiên cứu đối với từng quốc gia, tính tới sự khác biệt giữa nghiên cứu hàn lâm và phi hàn lâm.

Chỉ số #1 đo lường chi phí trực tiếp các nhà nghiên cứu và các nhà cấp vốn ngày nay đối mặt.

Chỉ số #2: Chi phí lưu trữ

Chi phí lưu trữ liên quan tới các chi phí lưu trữ điện tử các bản sao dư thừa thêm của dữ liệu mà nếu khác là không cần thiết nếu dữ liệu là FAIR. Tổng chi phí lưu trữ được ước tính dựa vào số lượng các bản sao, kích cỡ các tập hợp dữ liệu và thời gian lưu trữ.

Chỉ số #2 đo lường chi phí trực tiếp các tổ chức nghiên cứu ngày nay đối mặt.

Chỉ số #3: Chi phí cấp phép

Chỉ số này xem xét các chi phí cấp phép để sử dụng dữ liệu có thể được làm cho sẵn sàng như là dữ liệu mở hoặc dữ liệu FAIR.

Chỉ số #3 đo lường chi phí trực tiếp ngày nay các tổ chức và các nhà cấp vốn nghiên cứu phải trả.

Chỉ số #4: Rút lại nghiên cứu

Rút lại nghiên cứu đo lường chi phí nghiên cứu lẽ ra không bị rút lại nếu các nguyên tắc FAIR được tôn trọng.

Chỉ số #4 đo lường chi phí gián tiếp ngày nay các nhà nghiên cứu đối mặt.

Chỉ số #5: Cấp vốn hai lần

Những gì có thể đạt được khi đo lường, nghĩa là những lợi ích đối với lợi ích đan xen nhiều hơn, là không thể trực tiếp. Chúng tôi quy định rằng nghiên cứu không FAIR dẫn tới đúp bản các dự án nghiên cứu và các chi phí liên quan tới nghiên cứu dư thừa phản ánh ít nhất một phần giá trị của nghiên cứu, thay vào đó, có thể được cấp vốn. Chỉ số này không bao gồm sự đúp bản vài hoạt động nghiên cứu giữa các dự án nghiên cứu khác nhau. Chúng được đưa vào trong chỉ số #1: thời gian bỏ ra.

Chỉ số #5 đo lường chi phí trực tiếp ngày nay các nhà cấp vốn đối mặt.

Chỉ số #6: Liên ngành

Liên ngành tham chiếu tới giá trị gia tăng của nghiên cứu mới kết hợp vài ngành học thuật là có thể nhờ các nguyên tắc FAIR, so với giá trị của nghiên cứu có thể được thực hiện nếu không là FAIR.

Chỉ số #7: Tăng trưởng kinh tế tiềm năng

Tăng trưởng kinh tế tiềm năng tham chiếu tới tăng trưởng GDP và số lượng công ăn việc làm có thể được tạo ra nếu các nguyên tắc FAIR được áp dụng rộng rãi. Bằng việc mở khóa giá trị của nghiên cứu và tạo thuận lợi cho quy trình khoa học, FAIR có tác động tích cực lên đổi mới sáng tạo biến bản thân nó thành tạo ra công ăn việc làm và GDP cao hơn.

Chỉ số #7 ước lượng chi phí cơ hội vì những lợi ích của FAIR chưa được hiện thực hóa.

2.4 Định lượng các chỉ số

Để định lượng các chỉ số được lựa chọn, chúng tôi đã bắt đầu bằng việc thu thập dữ liệu ở mức độ kinh tế vi mô, ấy là dữ liệu áp dụng được cho các dự án và các hoạt động nghiên cứu riêng lẻ. Một khi chúng tôi đã định lượng được các chi phí không có dữ liệu FAIR ở mức vi mô, chúng tôi đã ngoại suy những kết quả đó đến cấp độ của nền kinh tế nghiên cứu châu Âu.

Vì việc ngoại suy từ mức vi mô sang vĩ mô, biên độ lỗi và khoảng ước lượng ở mức vi mô có thể làm cho ước lượng toàn cầu trở nên không chính xác. Vì thế, chúng tôi ước tính thận trọng nhất về khía cạnh các chi phí không có dữ liệu FAIR và tần suất ở đó các chi phí đó xảy ra. Các chỉ số không được định lượng rõ ràng sẽ bị loại trừ. Hệ quả là, chi phí đúng thực sự của việc không có dữ liệu FAIR có thể cao đáng kể, vì các hiệu ứng lan tỏa không thể định lượng hoặc các ngoại tác định tính vẫn chưa được tính đến.

Khi thu thập dữ liệu cho các chỉ số kinh tế vi mô, chúng tôi đã áp dụng các tiêu chí sau đây để tối đa hóa tính đại diện của dữ liệu:

- Đặc thù ngành: dữ liệu có liên quan tới ngành nghiên cứu đặc thù sẽ được ưu tiên hơn so với dữ liệu tổng hợp cho một chương trình nghiên cứu. Điều này sẽ cho phép chúng tôi ngoại suy thông tin chính xác hơn, vì các ngành khác nhau có trọng số khác nhau.
- Độ phủ địa lý: các tập hợp dữ liệu của chúng tôi bao trùm nhiều quốc gia (ví dụ, Đan Mạch, Phần Lan, Hà Lan, Vương quốc Anh), hầu hết ở châu Âu. Việc hiểu biết các chi tiêu cho nghiên cứu và các lĩnh vực nghiên cứu của từng quốc gia cho phép chúng tôi ngoại suy thông tin chính xác hơn. Nhưng vì thiếu các nguồn chất lượng, các kết quả từ các khu vực ngoài châu Âu (ví dụ, Úc) đã được đưa vào trong báo cáo của chúng tôi khi hạ tầng nghiên cứu là tương tự.
- Khả năng sử dụng lại: dữ liệu liên quan đến các trường hợp không có khả năng tái diễn đã được tránh đến mức có thể.

2.4.1 Chỉ số #1: Thời gian bỏ ra

Để ước lượng thời gian[47] bị lãng phí vào các hoạt động nghiên cứu nhất định, là cơ bản phải có ý tưởng gần đúng về thời gian các nhà nghiên cứu bỏ ra cho từng hoạt động nghiên cứu.

Chi phí thời gian bỏ ra đã được xác định bằng việc nhân thời gian lãng phí vì không có dữ liệu FAIR với lương trung bình của các nhà nghiên cứu hàn lâm và phi hàn lâm một cách tương ứng. Tính toán này sẽ tính tới số lượng các nhà nghiên cứu và lương trung bình cho từng trong số 28 quốc gia thành viên của Liên minh châu Âu.

$$Cost\ of\ time\ spent = Time_{Wasted} \times Wages$$

Thời gian lãng phí của các nhà nghiên cứu được xác định bằng cách tính thời gian các nhà nghiên cứu dành cho các hoạt động nghiên cứu[48] và sự thiếu hiệu quả do dữ liệu không là FAIR.

$$Time_{Wasted} = Time_{Activities} \times Inefficiency\ rate$$

Phần thời gian các nhà nghiên cứu bỏ ra cho các hoạt động nghiên cứu đã được (Ziker, 2013) và (Tenopir, et al., 2011) và những người khác nghiên cứu. Để tính toán, chúng tôi sử dụng giá trị trung bình tổng hợp từ bảy nghiên cứu, hai nghiên cứu được đề cập ở trên và (Nur Farah Naadia Mohd Fauzi, Abd Rashid, Ahmad Sharkawi, Farihah Hasan, & Aripin, 2016), (Court, 2012), (Cheol Shin, Arimoto, Cummings, & Teichler, 2014), (Houghton & Gruen, 2014). Ngoài ra, chúng tôi cũng phân tách giữa các bài báo khoa học (báo cáo) và dữ liệu hỗ trợ trong tính toán của chúng tôi bất cứ ở đâu có liên quan.

Nhìn chung, sự không hiệu quả về thời gian vì thiếu dữ liệu FAIR đã được ước lượng là 3,12% đối với các nhà nghiên cứu hàn lâm với phần thời gian bỏ ra cho từng trong số các hoạt động nghiên cứu sau đây và những điều không hiệu quả có liên quan. Biết được tỷ lệ thời gian dành cho nghiên cứu của các nhà nghiên cứu hàn lâm và phi hàn

47 For the remaining of this section, the term ‘Time’ will be used in equations as a percentage of time but for convenience, we will keep the notation ‘Time’.

48 The research activities are defined in section 2.2.1.

lâm[49], chúng tôi đã ngoại suy con số này thành 4,47% thời gian không hiệu quả đối với các nhà nghiên cứu phi hàn lâm.

2.4.1.1 Tạo lập và thu thập dữ liệu

Chúng tôi đã ước tính rằng 31,52% thời gian bỏ ra cho việc tìm kiếm dữ liệu có thể tiết kiệm được nếu các nguyên tắc FAIR được áp dụng. Số liệu này dựa vào các thống kê tổng hợp[50] về sử dụng và chất lượng của siêu dữ liệu (ví dụ, dạng các tiêu chuẩn, nếu có) và giả thiết về việc mất thời gian tùy thuộc vào tính sẵn sàng và chất lượng của siêu dữ liệu.

Thời gian tìm kiếm dữ liệu thứ cấp có liên quan trực tiếp tới chất lượng và sự phong phú của siêu dữ liệu sẵn có. Hệ quả là, chúng tôi đã ước tính thời gian tuân theo 4 mức về tính sẵn sàng của siêu dữ liệu:

- Không có siêu dữ liệu;
- Siêu dữ liệu không tuân thủ bất kỳ tiêu chuẩn nào;
- Siêu dữ liệu tuân thủ các đặc tả địa phương/sở hữu độc quyền; và
- Siêu dữ liệu tuân theo các tiêu chuẩn được quốc tế thừa nhận như DataCite[51], DCAT-AP[52] Dublin Core[53], DDI[54], hoặc SDMX[55].

Đối với thời gian để tìm kiếm dữ liệu thứ cấp đúng, chúng tôi bổ sung thời gian bị mất trong việc tìm kiếm dữ liệu ở nhiều điểm truy cập. Từng nhà nghiên cứu mất thời gian truy cập dữ liệu thích đáng từ các tạp chí nơi sự xác thực được yêu cầu. Thời gian bị mất này gồm thời gian được yêu cầu để xác thực với tạp chí đó cũng như thời gian cho việc quản lý một tài khoản (ví dụ như: tạo lập, xác thực, thay đổi mật khẩu). Chúng tôi giả thiết thời gian này là khá thấp. Tuy nhiên, việc triển khai các nguyên tắc FAIR có thể làm giảm lượng thời gian nhà nghiên cứu sẽ cần để xác thực bản thân mình.

49 (Tenopir, et al., 2011)
50 (Tenopir, et al., 2011) (Royal Veterinary College, University of London, n.d.) (Parsons, Grimshaw, & Williamson, 2013)
51 http://schema.datacite.org/
52 https://joinup.ec.europa.eu/page/dcat-ap
53 http://dublincore.org/specifications/
54 http://www.ddialliance.org/Specification/
55 https://sdmx.org/

2.4.1.2 Tiền xử lý và làm sạch dữ liệu

Trong khi các nguyên tắc FAIR về tính tương hợp và khả năng sử dụng lại có thể tạo thuận lợi cho việc (tiền) xử lý các dữ liệu hiện có, thời gian tiết kiệm được có liên quan trực tiếp tới các nguyên tắc FAIR chủ yếu vì làm giảm được nỗ lực cần thiết để ban đầu làm sạch dữ liệu.

Tuy nhiên, thời gian bỏ ra trong bước này thường khó để tách khỏi thời gian cho các hoạt động chuyển đổi và tích hợp. Chúng tôi vì thế đã ước tính chi phí thời gian cho bước này cùng với bước tích hợp dữ liệu.

Ngoài ra, chúng tôi hiểu từ các cuộc phỏng vấn với các nhà nghiên cứu và các chuyên gia rằng hầu hết thời gian FAIR tiết kiệm được thuộc về các hoạt động nghiên cứu khác và sự tác động lên sản xuất dữ liệu từ các nhà nghiên cứu là không đáng kể.

2.4.1.3 Tích hợp dữ liệu

Khả năng được phép sử dụng lại một bài báo và thời gian cần thiết để xác định quyền này tùy thuộc vào giấy phép được cung cấp và sự dễ dàng mà với nó một nhà nghiên cứu có thể tìm thấy giấy phép này. Từ đó, chúng tôi đã ước tính thời gian trung bình cần thiết để xác định và hiểu một giấy phép cho dữ liệu thứ cấp và đã ước tính rằng FAIR có thể giúp làm giảm thời gian này tới 1,46%.

Trong khi chúng tôi đã thấy rằng thời gian cần thiết để xác định một giấy phép là thấp, cần lưu ý là mỗi lần một tập hợp dữ liệu hoặc một tài liệu không thể được sử dụng lại, thì thời gian bỏ ra để tìm kiếm nó, về cơ bản, đã bị lãng phí.

Ngoài ra, có khả năng tích hợp nhiều nguồn dữ liệu thứ cấp với nhau trực tiếp tùy thuộc vào các định dạng và chất lượng dữ liệu sẽ được tích hợp và các mục tiêu nghiên cứu đó theo đuổi. Để ước lượng tính khả thi của sự tích hợp, chúng tôi xem xét tỷ lệ phần trăm dữ liệu nhân bản được, liệu định dạng đó của dữ liệu máy có đọc được hay không, và liệu các vấn đề về chất lượng dữ liệu có được giải quyết hay không.

- Dữ liệu nhân bản được:
    - Máy đọc được; hoặc
    - Máy không đọc được, về điều đó chúng tôi phân biệt định dạng PDF với các định dạng khác.

- Dữ liệu không nhân bản được:
  - Có khả năng giải quyết; hoặc
  - Không có khả năng giải quyết.

Đối với từng mức chất lượng dữ liệu ở trên, chúng tôi đã ước tính thời gian bị mất cho việc có khả năng sử dụng lại dữ liệu đó và làm thế nào FAIR có thể tác động tới thời gian này. Ví dụ đối với dữ liệu có khả năng giải quyết được với các vấn đề chất lượng, chúng tôi đã dựa bản thân vào một khảo sát từ (Biology, 2015) với thời gian cần thiết để sửa dữ liệu chất lượng kém.

Tác động của FAIR lên thời gian và khả năng tương tác dữ liệu với nhau là ở hai khía cạnh. Trước hết, FAIR thúc đẩy các nhà nghiên cứu sử dụng các định dạng dữ liệu là "chính thống, truy cập được, chia sẻ được, và áp dụng được rộng rãi để trình bày kiến thức"[56] cho người và máy đều đọc được. Khi nhiền báo cáo chỉ truy cập được ở định dạng PDF, việc máy đọc được bị hạn chế. Thứ hai, nguyên tắc FAIR sử dụng lại được nêu rằng dữ liệu phải "được mô tả phong phú với với nhiều thuộc tính chính xác và có liên quan"[57] cho phép người và máy "quyết định liệu dữ liệu đó có thực sự là hữu ích hay không trong một ngữ cảnh cụ thể"[58].

Tổng thể, chúng tôi đã ước tính rằng FAIR có thể giúp làm giảm thời gian bỏ ra để tích hợp dữ liệu tới 28,66%.

2.4.1.4 Phân tích dữ liệu

Tác động của việc không có FAIR lên thời gian bỏ ra để phân tích dữ liệu là khó tách bạch với thời gian bỏ ra vì các vấn đề về tính tương hợp và khả năng sử dụng lại phải đối mặt khi tích hợp dữ liệu.

Vì thiếu dữ liệu đặc thù về tác động của FAIR lên hoạt động này, chúng tôi đã ước tính thời gian bổ sung thêm cần thiết đặc biệt cho phân tích sẽ là đáng kể.

56 https://www.go-fair.org/fair-principles/i1-metadata-use-formal-accessible-shared-broadly-applicable-language-knowledge-representation/

57 https://www.go-fair.org/fair-principles/r1-metadata-richly-described-plurality-accurate-relevant-attributes/

58 https://www.go-fair.org/fair-principles/r1-metadata-richly-described-plurality-accurate-relevant-attributes/

2.4.1.5 Đăng ký và xuất bản

Khi nói về đăng ký và xuất bản, chúng tôi thấy rằng tác động chính của FAIR không phải là lên thời gian bỏ ra cho việc đăng ký và xuất bản dữ liệu, mà lên số lượng các bản sao dư thừa đang được làm ra vì dữ liệu dữ liệu không truy cập được. Các khía cạnh đó được định lượng bằng chỉ số #2: Chi phí lưu trữ.

2.4.1.6 Rà soát lại ngang hàng

Chúng tôi đã giả thiết rằng 10% thời gian bỏ ra cho việc rà soát lại ngang hàng có thể tiết kiệm được nếu các nguyên tắc FAIR được áp dụng. FAIR tác động tới rà soát lại ngang hàng bằng việc nâng cao chất lượng của nghiên cứu và khả năng tái tạo lại các kết quả. Mặc dù chúng tôi cho rằng nhu cầu rà soát lại ngang hàng sẽ vẫn chủ yếu, nhưng FAIR sẽ giúp làm giảm thời gian cần thiết để thực hiện rà soát lại đó.

Sau đó, chúng tôi tính đến giả định này với lượng rà soát lại ngang hàng và lượng thời gian dành cho rà soát lại ngang hàng từ các tài liệu hiện có để xác định chi phí của việc không có FAIR.

Các nguyên tắc FAIR có thể có tác động phụ lên thời gian các nhà nghiên cứu bỏ ra cho việc rà soát lại ngang hàng. Mỗi khi một báo cáo đã xuất bản bị rút lại khỏi một tạp chí vì những lý do như sai sót trong phương pháp luận hoặc hành vi sai trái của tác giả, thì thời gian mà những người rà soát lại ngang hàng bên ngoài đã dành cho bài báo này bị lãng phí. Vì các nguyên tắc FAIR có thể có tác động lên số lượng các báo cáo có khả năng bị rút lại như được mô tả trong chỉ số #3: Rút lại nghiên cứu, nó cũng có thể đóng góp vào việc làm giảm thời gian lãng phí đối với những người rà soát lại ngang hàng.

2.4.2 Chỉ số #2: Chí phí lưu trữ

Đối với các nhà xuất bản và các kho dữ liệu, các nguyên tắc FAIR có thể làm giảm chi phí lưu trữ dữ liệu bằng việc làm giảm nhu cầu đối với các bản sao dư thừa. Để định lượng chi phí này, chúng tôi sử dụng các thông tin sau:

- Lượng dữ liệu nghiên cứu đối với từng nhà nghiên cứu châu Âu mỗi năm[59].
- Tỷ lệ các tập hợp dữ liệu được lưu trữ ở nhiều nơi[60].

59 Based on sources such as (Addis, 2015) and (Van Tuyl & Michalek, 2015)

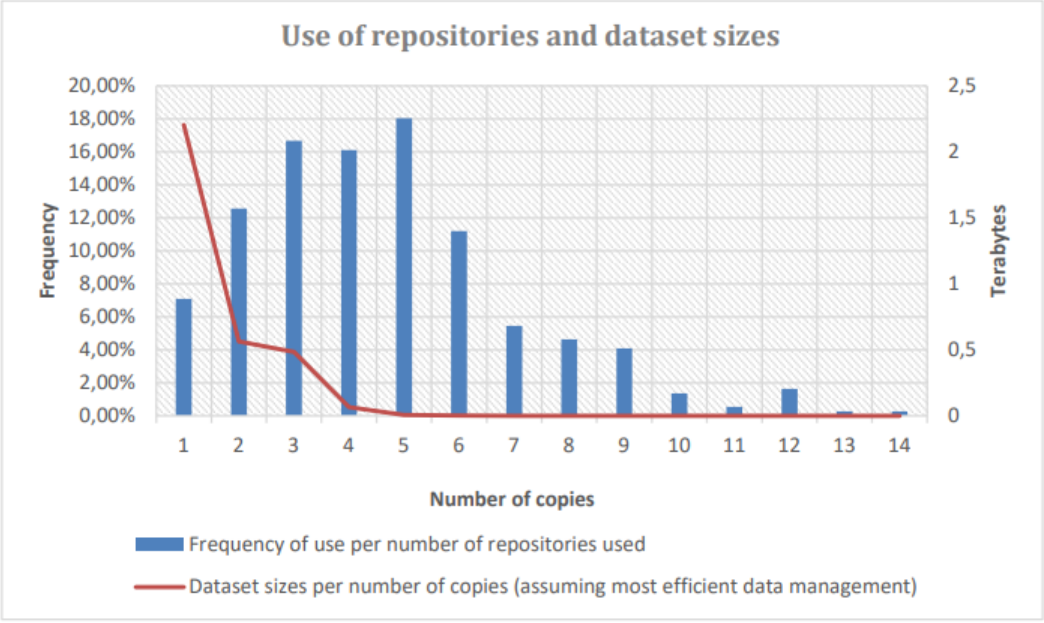
60 (Parsons, Grimshaw, & Williamson, 2013)

- Chi phí lưu trữ trên đám mây và ở trường đại học.

Dựa vào các mô hình định giá khác nhau[61], chúng tôi đã lưu ý rằng chi phí lưu trữ dữ liệu thường trực có quan hệ tuyến tính với lượng dữ liệu được lưu trữ.

Các giả thiết sau đây[62] đã được sử dụng để định lượng chi phí lưu trữ bổ sung vì dữ liệu không FAIR:

- Bằng chứng thực nghiệm[63] cho thấy trung bình dữ liệu được lưu trữ trong 4,63 kho. Trong một thế giới mở và cộng tác hoàn hảo, dữ liệu có lẽ không cần phải được lưu trữ trong nhiều hơn 1 kho. Tuy nhiên, FAIR không là thuốc chữa bách bệnh cho sự dư thừa không cần thiết. Để loại bỏ dư thừa không cần thiết, các quy định nội bộ và các thay đổi văn hóa cũng là cần thiết trong các cơ sở (cấp vốn) nghiên cứu. Trong khi FAIR có thể đóng góp để làm cho sự thay đổi này xảy ra, chúng tôi giả thiết bảo thủ rằng việc triển khai các nguyên tắc FAIR sẽ trực tiếp làm giảm số lượng các bản sao dữ liệu dư thừa tới 20%. Đồ thị bên dưới chỉ ra sự phân bổ số lượng các bản sao (ví dụ, các kho được sử dụng) màu xanh da trời và kích cỡ trung bình của các tập hợp dữ liệu theo số lượng các bản sao.



Hình 4: Phân bổ lượng các bản sao theo số lượng các kho

61 https://aws.amazon.com/s3/pricing/ and (Royal Veterinary College, University of London, n.d.)
62 Arkivum, Estimating Research Data Volumes in UK HEI, 2015.
63 (Parsons, Grimshaw, & Williamson, 2013)

Nếu chúng tôi giả thiết rằng các tập hợp dữ liệu lớn nhất được quản lý tốt hơn (vì các chi phí lưu trữ cao hơn) và bản đồ phân bổ như được trình bày ở trên[64], thì việc có một bản sao dữ liệu duy nhất có thể làm giảm tổng lượng được lưu trữ đến 25,67%. Điều này ngụ ý là giảm 20% trên diện rộng vừa khả thi vừa bảo toàn vì trong thực tế, không phải tất cả các tập hợp dữ liệu lớn nhất sẽ được lưu giữ trong một kho lưu trữ duy nhất.

- Sao lưu được thực hiện như một phần của việc vận hành một hạ tầng hoặc một kho không được coi là các bản sao tách biệt khỏi dữ liệu. Đối với tính toán của chúng tôi, chúng tôi coi các bản sao lưu như một biện pháp kỹ thuật để ngăn ngừa mất dữ liệu đối với một bản sao dữ liệu.
- Vì chúng tôi dựa nhiều vào các khảo sát và dữ liệu từ Vương quốc Anh (theo tư liệu, Vương quốc Anh là một trong các quốc gia tiên tiến nhất về quản lý dữ liệu) chúng tôi giả thiết rằng dữ liệu các giả thiết và tính toán chúng tôi dựa vào có thể được ngoại suy cho 28 quốc gia của Liên minh châu Âu;
- Chúng tôi hiểu lưu trữ thường trực tối thiểu là 20 năm;
- Tổng lượng dữ liệu được ước lượng dựa vào các nghiên cứu hiện hành sử dụng các đơn vị dữ liệu để hỏi các nhà nghiên cứu về lượng dữ liệu được sử dụng. Các đơn vị dữ liệu đó không có phân bổ tuyến tính, ví dụ: 1-50MB, 50-100MB, 100-250MB, .v.v[65]. Hệ quả là, phép ngoại suy được thực hiện từ các khảo sát như vậy được lưu giữ trong mỗi đơn vị dữ liệu được sử dụng đó;
- Bên cạnh sự phân bổ phi tuyến tính, đề xuất cuối cùng của các đơn vị dữ liệu đó thường là kết thúc mở, nghĩa là lớn hơn 1TB. Điều này có thể tạo ra khuynh hướng trong tính toán của chúng tôi cho việc lưu trữ các tệp lớn nhất, vì một câu trả lời trong chủng loại này có thể là 2TB hoặc 10TB, với ảnh hưởng đáng kể lên giá trị trung bình.

Chi phí định giá theo lượng dữ liệu dao động với thời gian và biến động theo mô hình định giá của kho. Trong nghiên cứu này, chúng tôi sử dụng mô hình định giá được một trường cao đẳng của Vương quốc Anh chào cho các nhà nghiên cứu của riêng nó[66], và

64 This assumption also gives the most conservative figures.
65 (Van Tuyl & Michalek, 2015) and (Addis, 2015)
66 (Royal Veterinary College, University of London, n.d.)

các mô hình định giá lưu trữ dữ liệu của AWS[67], Azure[68] và Google[69]. Cuối cùng, vì tính sẵn sàng hạn chế của các số liệu về lượng dữ liệu cho từng nhà nghiên cứu và sự dư thừa, mục đích tính toán của chúng tôi bị giới hạn cho nghiên cứu học thuật. Việc mở rộng các kết quả của chúng tôi cho toàn bộ khu vực nghiên cứu ở Liên minh châu Âu có thể giả thiết rằng tất cả các nhà nghiên cứu ở châu Âu sản xuất lượng dữ liệu y hệt nhau mỗi năm.

2.4.3 Chỉ số #3: Các chi phí cấp phép

Trước nhất, chúng tôi đã tính toán tỷ lệ phần trăm của các báo cáo chỉ ra một giấy phép sử dụng được trong siêu dữ liệu của chúng và chúng tôi đã xem xét các nơi ở đó giấy phép đó thực sự đã được chỉ định. Việc sử dụng các giấy phép khác nhau đã được định giá:

- liệu có việc cấp phép không, nhà nghiên cứu hoặc:
  - Không sử dụng lại báo cáo và dữ liệu của nó; hoặc
  - Sử dụng lại báo cáo và dữ liệu của nó mà không biết liệu anh/chị ta có thể hay không.
- Liệu có một giấy phép trong siêu dữ liệu của báo cáo hoặc của dữ liệu mà có thể:
  - Là giấy phép được tiêu chuẩn hóa máy đọc được; hoặc
  - Một giấy phép địa phương người đọc được; và
  - Một giấy phép mở; hoặc
  - Một giấy phép đóng áp đặt một khoản phí để sử dụng lại dữ liệu đó.

Để định lượng các chi phí cấp phép có liên quan tới việc không có FAIR và không có truy cập mở, chúng tôi đã tập hợp các số liệu thống kê về phân vùng sử dụng và chúng tôi đã xem xét tỷ lệ phần trăm dữ liệu nào có thể được làm thành mở. Bằng việc tổng hợp các nghiên cứu[70] về truy cập mở tới dữ liệu, chúng tôi thấy rằng 71,5% dữ liệu có

67 https://aws.amazon.com/pricing/
68 https://azure.microsoft.com/en-us/pricing/calculator/#storage1
69 https://cloud.google.com/products/calculator/
70 (Tenopir, et al., 2011) and (Johnson, Parsons, Chiarelli, & Kaye, JISC Research Data Assessment Support - Findings of the 2016 data assessment framework (DAF) surveys, 2016)

thể được làm thành mở. Điều này khiến 28,5% dữ liệu nghiên cứu học thuật vẫn phải được đóng lại vì lý do bảo mật và quyền riêng tư.

Thứ hai, chúng tôi đã tính đến tỷ lệ phần trăm dữ liệu đã được mở rồi ngày nay. Theo Ủy ban châu Âu[71], trong năm 2015, 48% dữ liệu là sẵn sàng theo truy cập mở (nghĩa là truy cập mở vàng và xanh). Tương tự một nghiên cứu gần đây đã ước lượng, cũng cho năm 2015, là 45% dữ liệu là sẵn sàng theo truy cập mở.

Thứ ba, chúng tôi đã xem xét các chi phí thuê bao và truy cập đối với các cơ sở nghiên cứu. Chúng tôi thấy dữ liệu đối với các tổ chức nghiên cứu công lập ở Vương quốc Anh[72] và Phần Lan[73] mà chúng tôi đã sử dụng để ngoại suy các chi phí của các giấy phép cho các nhà nghiên cứu hàn lâm ở 28 quốc gia Liên minh châu Âu dựa vào số lượng các nhà nghiên cứu ở từng quốc gia. Chỉ số này chỉ xem xét các chi phí cấp phép phát sinh ngày nay. Nó không bao gồm các chi phí cấp phép giả định cho dữ liệu lẽ ra đã được sử dụng lại nếu nó đã sẵn sàng rồi theo truy cập mở.

2.4.4 Chỉ số #4: Rút lại nghiên cứu

Để định lượng chi phí rút lại các bài báo, chúng tôi trước nhất đã xác định số lượng các lý do cho việc rút lại nghiên cứu. Các nghiên cứu khác nhau[74] ước tính các lý do (ví dụ, không có khả năng tái tạo lại, lỗi, giả mạo, đạo văn, .v.v.) và số lượng các bài báo bị rút lại (0,04% tất cả các bài báo). Dựa vào các thảo luận với các chuyên gia và các tính toán, chúng tôi giả thiết rằng FAIR có thể giúp làm giảm lượng các bài báo bị rút lại tới 51,44%. Tỷ lệ phần trăm này là dựa vào nghiên cứu[75] xem xét các lý do rút lại nghiên cứu và các giả định của chúng tôi về tác động của FAIR lên từng trong số các lý do đó (ví dụ, lỗi, không có khả năng tái tạo lại, giả mạo (nghi ngờ), gửi nhiều lần, đạo văn, các hành xử không đúng khác). Chúng tôi sau đó tính thời gian các nhà nghiên cứu bỏ ra cho nghiên cứu bị rút lại và xác định chi phí của nó.

71 https://ec.europa.eu/research/openscience/index.cfm?pg=access&section=monitor#viz1489066430689

72 (Lawson, Meghreblian, & Brook, 2015)

73 (Lahti, 2016)

74 (Horbach & Halffman, 2017), (Masic, 2012), (Fang, Steen, & Casadevall, 2012), and (Wager & Williams, 2011)

75 (Wager & Williams, 2011)

Ngoài việc làm giảm tỷ lệ các bài báo bị rút lại, các nguyên tắc FAIR còn có thể tác động lên thời gian trung bình cần thiết trước khi một bài báo bị rút lại. Tuy nhiên, vì chi phí có liên quan tới thời gian trung bình rút lại là khó ước tính, chúng tôi đã không định lượng nó.

2.4.5 Chỉ số #5: Cấp vốn hai lần

Các ước tính của chúng tôi về việc cấp vốn hai lần tập trung vào các chương trình nghiên cứu và các trợ cấp nghiên cứu khu vực công. Để xác định ở mức độ nào việc cấp vốn hai lần xảy ra, chúng tôi dựa vào tác phẩm của Harold R. Garner, Lauren J. McIver và Michael B. Waitzkin[76], nó ước lượng việc cấp vốn hai lần bằng việc sử dụng các thuật toán toàn văn để xác định sự chồng chéo trong các đơn xin trợ cấp. Từ tác phẩm của họ, chúng tôi đã sử dụng tỷ lệ phần trăm các đơn xin trợ cấp với sự chồng chéo đáng ngờ và tỷ lệ miêu tả kích cỡ trung bình trao trợ cấp đầu tiên được so sánh với sự trao trợ cấp tiềm tàng có sự chồng chéo.

Các kết luận của họ cho các chương trình nghiên cứu của Mỹ được ngoại suy cho các chương trình nghiên cứu của châu Âu bằng việc sử dụng sự đóng góp trung bình của Liên minh châu Âu cho các dự án được H2020 cấp vốn cùng với tổng số trợ cấp nghiên cứu công ở Liên minh châu Âu.

Việc ngăn ngừa cấp vốn hai lần chủ yếu dựa vào các công cụ chống đạo văn khi tự động so sánh nghiên cứu với các bài báo và dữ liệu đang có. Phương pháp này chỉ hiệu quả ở mức độ nghiên cứu có đạo văn là sẵn sàng ở định dạng máy đọc được, điều có thể được đảm bảo bằng việc ứng dụng các nguyên tắc FAIR. Vì thế, chúng tôi coi rằng ít nhất 80% việc cấp vốn hai lần có thể tránh được với FAIR, vì ít có khả năng thấy các phương pháp sáng tạo mới nào cho đạo văn.

Cuối cùng, trong khi vài tổ chức cấp vốn nghiên cứu của tư nhân cũng sử dụng các trợ cấp, không có dữ liệu nào có thể được sử dụng để hỗ trợ cho các tính toán đó. Hệ quả là, các ước tính của chúng tôi đề cập tới việc cấp vốn hai lần trong nghiên cứu được nhà nước cấp vốn.

76 (Garner, McIver, & Waitzkin, 2013)

2.4.6 Chỉ số #6: Liên ngành

Để định lượng liên ngành, chúng tôi cần xác định lượng nghiên cứu được làm cho có thể thông qua lợi ích đan xen được FAIR xúc tác. Tuy nhiên, có ít nghiên cứu có ý định để định lượng tác động của liên ngành và các nghiên cứu mà thực hiện việc đó dựa vào các trường hợp điển hình[77]. Vì lý do này, tác động của FAIR lên liên ngành không thể không được ước tính với dữ liệu hiện có.

Tuy nhiên, mặc dù không có dữ liệu để ước tính giá trị của lợi ích đan xen, nhưng cũng cần lưu ý rằng một dự án nghiên cứu mới được thực hiện với FAIR sẽ luôn thay thế một dự án nghiên cứu khác đã diễn ra theo cách khác. Tác động của FAIR lên lợi ích đan xen, trên thực tế, không làm gia tăng tổng lượng nghiên cứu được thực hiện.

Do đó, tác động của FAIR đối với tính liên ngành chỉ giới hạn ở giá trị gia tăng của nghiên cứu mới so với giá trị của nghiên cứu sẽ được thực hiện nếu không có FAIR.

2.4.7 Chỉ số #7: Tăng trưởng kinh tế tiềm năng

Để định lượng tăng trưởng kinh tế tiềm năng, vài cách tiếp cận đã được bám theo, dựa vào tác phẩm của Beagrie & Houghton (2014, 2016) và cả tác phẩm của Tauri Group[78]. Do đó, chúng tôi đã sử dụng các quan điểm sau để xem xét tác động của dữ liệu nghiên cứu FAIR lên đổi mới sáng tạo, điều mà sau đó có thể được chuyển đổi thành một chi phí cơ hội:

- Sản phẩm phụ (Spinoff): dựa vào công việc của nhà thầu NASA, chúng tôi đã ước lượng tác động lên đổi mới sáng tạo thông qua mạng các sản phẩm phụ của châu Âu, vì các sản phẩm phụ bắt nguồn từ thế giới hàn lâm và dựa rất nhiều vào dữ liệu nghiên cứu. Chúng tôi đã giả thiết rằng vì dữ liệu nghiên cứu không là FAIR, một phần nhất định các sản phẩm phụ tiềm tàng không tồn tại. Tuy nhiên, hiệu ứng nhân lên của FAIR không thể định lượng được.
- Tác động hiệu quả: chúng tôi đã ước tính tác động hiệu quả đạt được do dữ liệu nghiên cứu FAIR. Dựa vào các tính toán trong phần 2.4.1, lãng phí thời gian vì không hiệu quả bắt nguồn từ dữ liệu không FAIR có thể được coi là thời gian được giải phóng và tái đầu tư để thực hiện các hoạt động liên quan đến nghiên

77 E.g. (Björkdahl, 2009)

78 https://www.nasa.gov/sites/default/files/files/SEINSI.pdf

cứu khác. Nói cách khác, vì dữ liệu nghiên cứu FAIR (nghĩa là truy cập tới các tài nguyên mới) một nhà nghiên cứu có thể sản xuất đầu vào mức y hệt với chi phí thấp hơn, điều được coi là có hiệu quả. Dù vậy, chúng tôi có thể ước tính với mức độ nào thời gian tiết kiệm được do FAIR có thể được tái đầu tư vào nghiên cứu mới đổi mới sáng tạo và giá trị của chúng.

- Định giá ngẫu nhiên: định giá ngẫu nhiên bao gồm ước tính giá trị của hàng hóa và dịch vụ phi thị trường dựa trên lý thuyết sở thích[79]. Trong trường hợp này, các cá nhân được yêu cầu những gì họ có thể trả tiền cho một hàng hóa hoặc dịch vụ trong một tình huống thị trường giả định (tức là trả tiền cho thứ gì đó thực sự miễn phí). Theo logic này, quy trình lý tưởng để định lượng dữ liệu nghiên cứu FAIR lên đổi mới sáng tạo có thể là quản lý một bảng câu hỏi với một kịch bản giả định. Bảng câu hỏi đó có thể được tạo ra theo một cách thức để đi tới một hệ số mô tả khả năng theo đó dữ liệu nghiên cứu FAIR có thể đóng góp cho công việc của một nhà nghiên cứu. Sau đó, hệ số đó có thể được chuyển đổi thành thiện chí trả tiền để hưởng lợi từ dữ liệu nghiên cứu FAIR[80]. Tuy nhiên, việc tiến hành khảo sát tăng cường để thu thập thông tin này từng là không thể trong khung thời gian và bối cảnh nghiên cứu của chúng tôi, điều trước hết dựa vào tư liệu hiện có.
- Mất dữ liệu: chúng tôi đã cố gắng xác định lượng dữ liệu bị mất vì không có dữ liệu nghiên cứu FAIR và ước tính giá trị của phần còn sót lại của dữ liệu đã bị mất. Tuy nhiên, không có thông tin khi nào và làm thế nào nhiều dữ liệu bị mất có liên quan tới việc không có FAIR, chúng tôi không thể ước tính được chi phí dữ liệu bị mất. Một vấn đề khác là không phải tất cả dữ liệu giữ được giá trị ngang bằng như nhau qua thời gian. Ví dụ, thông tin lưu thông và dự báo thời tiết mất nhiều nhất giá trị của chúng rất nhanh, trong khi dữ liệu khảo cổ học giữ được giá trị dài lâu sau phát hiện.
- Hoàn vốn đầu tư công: cuối cùng, chúng tôi cũng đã cố gắng đo lường hoàn vốn đầu tư công vào nghiên cứu FAIR, so sánh với nghiên cứu thông thường. Theo Houghton, hoàn vốn đầu tư công của xã hội vào nghiên cứu và phát triển (R&D)

79 Reference theory states that a good or service which contributes to human welfare has economic value.

80 To be also included, the calculation of the cost of (re)creating the data.

có thể đâu đó 20-60%. Tuy nhiên các vấn đề khác nhau phát sinh, trước nhất nghiên cứu vừa được cấp vốn công và tư, và sự phân bổ là không rõ. Và thứ hai là, tỷ lệ nào của hoàn vốn đầu tư công mà FAIR đang đóng góp.

Mối liên kết giữa đổi mới sáng tạo và tăng trưởng đã được các nhà kinh tế học và các nhà nghiên cứu khác nghiên cứu kỹ càng. Tuy nhiên, các mối liên kết như vậy luôn được minh họa với các trường hợp điển hình cụ thể, từ đó là không thể ngoại suy. Ví dụ, một trường hợp điển hình ở mức cơ sở có thể phải được ngoại suy ở mức quốc gia và châu Âu. Điều đó không bao gồm thực tế là một trường hợp điển hình nói chung là một chuyên ngành cụ thể, và do đó không thể phản ánh đúng thực tế của một chuyên ngành khác.

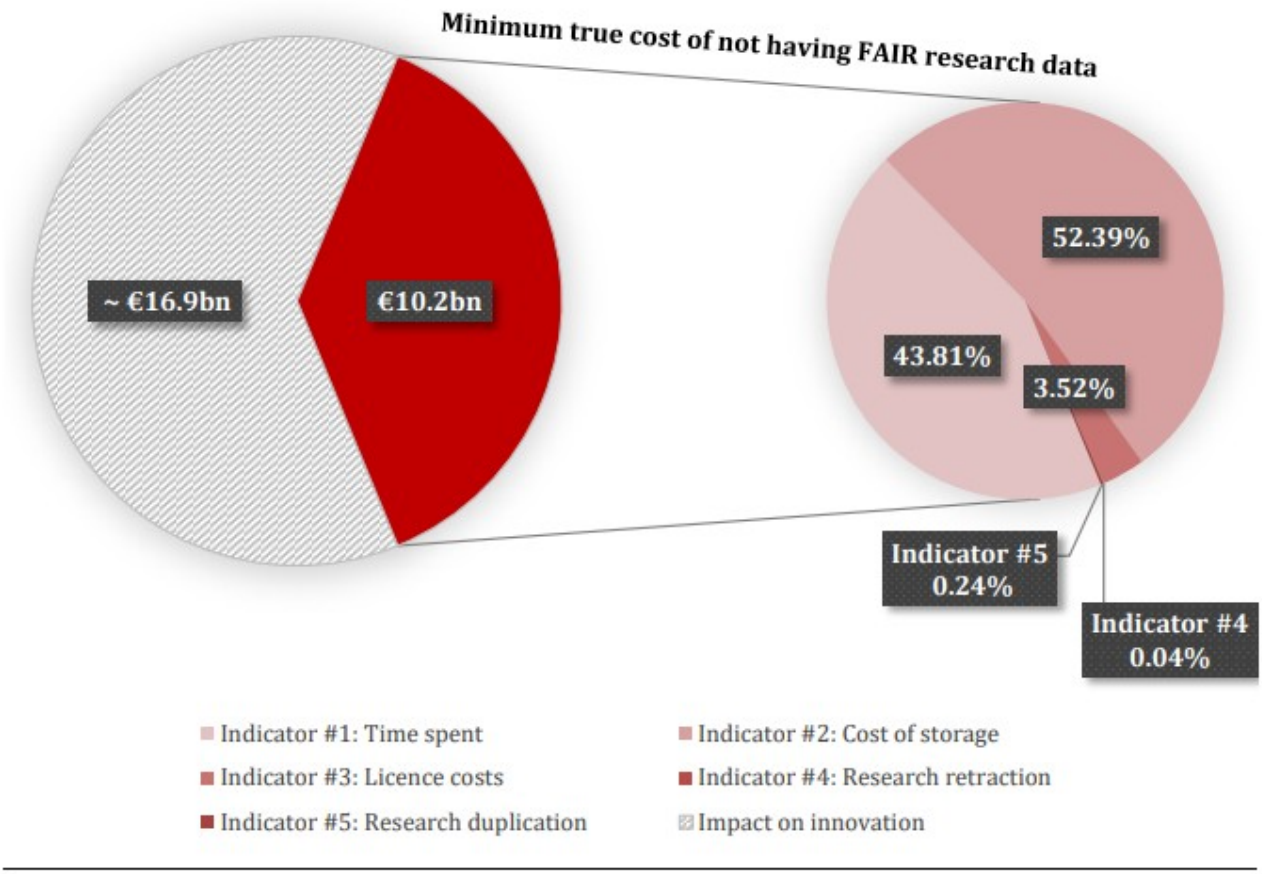
Ngoài ra, tác động của việc không có FAIR lên tăng trưởng kinh tế chỉ có thể được định lượng nếu tác động của FAIR lên đổi mới sáng tạo (ví dụ, các dự án hoặc các bằng sáng chế mới về khoa học, nghiên cứu) có thể được định lượng chính xác và không có chồng chéo. Tại thời điểm viết tài liệu này, không có đủ dữ liệu để hỗ trợ cho các giả thiết cần thiết để định lượng những lợi ích kinh tế của FAIR.

Thay vào đó, chúng tôi sẽ trình bày hầu hết các phát hiện định tính với khía cạnh về những lợi ích kinh tế của FAIR.

3. TÍNH TOÁN CHI PHÍ

Chúng tôi ước tính chi phí thường niên của việc không có dữ liệu FAIR tối thiểu là 10,2 tỷ € mỗi năm. Chi phí thực có khả năng còn cao hơn nhiều vì các yếu tố không định lượng được như giá trị của chất lượng nghiên cứu được cải thiện và các hiệu ứng lan tỏa tích cực gián tiếp khác của dữ liệu nghiên cứu FAIR.

Bức tranh sau đây chỉ ra sự phân bổ lại chi phí cho từng chỉ số được so sánh với tổng chi phí có khả năng của việc không có dữ liệu nghiên cứu FAIR, điều bao gồm sự ước tính dựa vào các số liệu cho dữ liệu mở. Ở phía bên trái của đồ thị, tác động lên đổi mới sáng tạo, có thể chiếm hơn 60% chi phí có khả năng vì không có dữ liệu nghiên cứu FAIR, trong khi chi phí tối thiểu thực sự của việc không có dữ liệu nghiên cứu FAIR, xoay quanh các chỉ số #1 tới #5 chiếm phần còn lại 40%.



Hình 5: Phân bổ chi phí

Phần bên phải của đồ thị chỉ ra rằng đối với chi phí thực tối thiểu của việc không có dữ liệu nghiên cứu FAIR, chỉ số #1: thời gian bỏ ra và chỉ số #2: chi phí lưu trữ chiếm hầu hết chi phí thực tối thiểu của việc không có dữ liệu nghiên cứu FAIR.

Các phần tiếp sau đây trình bày dữ liệu đầu vào chính và các giả thiết được sử dụng để tính toán cho từng chỉ số. Các chỉ số và cách tiếp cận để định lượng chúng được mô tả trong chương 2. Các tính toán của chúng tôi dựa hoàn toàn vào dữ liệu thứ cấp, là sẵn sàng công khai[81].

3.1 Chỉ số #1: Thời gian bỏ ra

Chí phí của thời gian bỏ ra vì không có nghiên cứu FAIR là 4,5 tỷ € mỗi năm. Cần lưu ý là việc có dữ liệu chỉ riêng đối với khu vực công, tác động của việc không có FAIR lên các nhân viên phi hàn lâm được ngoại suy từ tác động của việc không có FAIR lên các nhân viên hàn lâm. Bảng bên dưới trình bày dữ liệu đã được sử dụng để ước tính chỉ số này.

Chi phí mất thời gian đối với các nhân viên phi hàn lâm	***Đơn vị***	***2017***
Lương trung bình của nhà nghiên cứu khu vực tư nhân/nhân viên phi hàn lâm	EUR (€)	61.864€
Lương trung bình của nhà nghiên cứu trong chính phủ/nhân viên phi hàn lâm	EUR (€)	52.853€
Thời gian dành cho nghiên cứu đối với nhân viên phi hàn lâm	Thời gian	50,00%
Tác động của việc không có FAIR lên các nhân viên phi hàn lâm	Thời gian	4-47%
Số lượng các nhà nghiên cứu ở khu vực tư nhân/nhân viên phi hàn lâm	#	923.997
Số lượng các nhà nghiên cứu trong chính phủ/nhân viên phi hàn lâm	#	269.963
Tổng chi phí thời gian bị mất đối với các nhân viên phi hàn lâm		***3.221.109.809€***

Chi phí mất thời gian đối với các nhân viên hàn lâm	***Đơn vị***	***2017***
Lương trung bình của nhà nghiên cứu ở giáo dục đại học/nhân viên hàn lâm	EUR (€)	54.484,1€
Thời gian dành cho nghiên cứu đối với nhân viên hàn lâm	Thời gian	34,96%
Tác động của việc không có FAIR lên các nhân viên hàn lâm	Thời gian	3%
Số lượng các nhà nghiên cứu trong giáo dục đại học	#	727.348
Tổng chi phí thời gian bị mất đối với các nhân viên hàn lâm		***1.238.244.725€***
Tổng chi phí thời gian bị mất vì không có dữ liệu nghiên cứu FAIR		***4.459.354.534€***

Bảng 2: Tính toán chỉ số #1

81 https://ec.europa.eu/info/files/cost-not-having-fair-research-data-data-and-calculations_en

3.2 Chỉ số #2: Chi phí lưu trữ

Chi phí lưu trữ dư thừa vì không có nghiên cứu FAIR là 5,3 tỷ € mỗi năm. Bảng bên dưới trình bày dữ liệu đã được sử dụng để ước tính chi phí lưu trữ có thể không cần thiết nếu các nguyên tắc FAIR đã được áp dụng.

Đầu vào về sử dụng dữ liệu	***Đơn vị***	***Giá trị***
Lượng dữ liệu trung bình được 1 nhà nghiên cứu tạo ra trong 1 năm	TB	2,45
Chi phí trung bình để lưu trữ dữ liệu cho mỗi TB/năm	EUR (€)	122€
Số lượng trung bình các kho nơi dữ liệu được lưu trữ	#	4,63
Thời gian trung bình dữ liệu được giữ lại	Năm	10%
Giảm số lượng các kho cần thiết nhờ có FAIR	%	20%
Số lượng trung bình các kho nơi dữ liệu được lưu trữ sau khi triển khai FAIR	#	3,70
Kết quả: Chi phí lưu trữ đối với mỗi nhà nghiên cứu	EUR (€)	2.776
Đầu vào về số lượng các nhà nghiên cứu		**2017**
Tổng số các nhà nghiên cứu ở 28 quốc gia thành viên Liên minh châu Âu		***1.921.308***
Tổng chi phí lưu trữ không cần thiết vì không có dữ liệu nghiên cứu FAIR		***5.333.228.576€***

Bảng 3: Tính toán chỉ số #2

Cần lưu ý là lượng dữ liệu trung bình được mỗi nhà nghiên cứu tạo ra trong một năm xoay quanh tất cả các dạng dữ liệu (như, bản ghi âm và video, các hình ảnh, các tài liệu, các bảng tính, đánh dấu/mã, .v.v.).

3.3 Chỉ số #3: Các chi phí cấp phép

Chi phí của các chi phí cấp phép vì thiếu truy cập mở là 360 triệu € mỗi năm. Bảng bên dưới trình bày dữ liệu đã được sử dụng để ước lượng giá trị chi phí nảy sinh vì việc không có truy cập mở đầy đủ tới dữ liệu nghiên cứu.

Đầu vào về truy cập mở	***Đơn vị***	***2017***
Dữ liệu (nghiên cứu) hiện đang theo truy cập mở	%	56,09%
Dữ liệu (nghiên cứu) hiện không sẵn sàng theo truy cập mở	%	43,91%
Tổng tỷ lệ dữ liệu (nghiên cứu) có thể được làm thành mở	%	71,47%
Dữ liệu (nghiên cứu) đóng, không mở vì lý do chấp nhận được (quyền riêng tư)	%	28,53%
Dữ liệu bổ sung có thể làm thành mở vì áp dụng FAIR	%	31,38%
Các chi phí cấp phép đối với các tổ chức nghiên cứu hàn lâm ở EU28	EUR (€)	1.141.161.883€
Tổng chi phí cấp phép vì không có truy cập mở		***358.095.416€***

Bảng 4: Tính toán chỉ số #3

Vì thiếu dữ liệu, chúng tôi chỉ có thể ước tính các chi phí cấp phép các nhà nghiên cứu đối mặt trong khu vực công. Các chi phí cấp phép thực sự vì không có FAIR và dữ liệu nghiên cứu mở vì thế có khả năng thậm chí còn cao hơn.

3.4 Chỉ số #4: Rút lại nghiên cứu

Chi phí rút lại nghiên cứu vì nghiên cứu không FAIR là 4,4 triệu € mỗi năm. Bảng bên dưới trình bày dữ liệu đã được sử dụng để ước tính giá trị thời gian bị mất vì rút lại nghiên cứu mà có thể đã tránh được với FAIR.

Chi phí mất thời gian đối với các nhân viên phi hàn lâm	***Đơn vị***	***2017***
Lương trung bình của nhà nghiên cứu khu vực tư nhân/nhân viên phi hàn lâm	EUR (€)	61.864€
Lương trung bình của nhà nghiên cứu trong chính phủ/nhân viên phi hàn lâm	EUR (€)	52.853€
Thời gian dành cho nghiên cứu đối với nhân viên phi hàn lâm	Thời gian	50,00%
Thời gian bị mất vì rút lại nghiên cứu	Thời gian	0,0085%
Giảm các bài báo bị rút lại bằng FAIR	%	51,44%
Tác động của việc không có FAIR lên các nhân viên phi hàn lâm	Thời gian	0,0044%
Số lượng các nhà nghiên cứu ở khu vực tư nhân/nhân viên phi hàn lâm	#	923.997
Số lượng các nhà nghiên cứu trong chính phủ/nhân viên phi hàn lâm	#	269.963
Tổng chi phí thời gian bị mất đối với các nhân viên phi hàn lâm		***3.144.844€***

Chi phí mất thời gian đối với các nhân viên hàn lâm	***Đơn vị***	***2017***
Lương trung bình của nhà nghiên cứu ở giáo dục đại học/nhân viên hàn lâm	EUR (€)	54.484,1€
Thời gian dành cho nghiên cứu đối với nhân viên hàn lâm	Thời gian	34,96%
Thời gian bị mất vì rút lại nghiên cứu	Thời gian	0,0059%
Giảm các bài báo bị rút lại bằng FAIR	%	51,44%
Tác động của việc không có FAIR lên các nhân viên hàn lâm	Thời gian	0,0031%
Số lượng các nhà nghiên cứu trong giáo dục đại học	#	727.348
Tổng chi phí thời gian bị mất đối với các nhân viên hàn lâm		***1.208.927€***
Tổng chi phí thời gian bị mất vì không có dữ liệu nghiên cứu FAIR		***4.353.772€***

Bảng 5: Tính toán chỉ số #4

Cần lưu ý là FAIR cũng làm lợi cho chất lượng các bài báo không bị rút lại. Các lợi ích đó có thể không ước tính được nhưng nó có thể được lập luận rằng các bài báo bị rút lại chỉ là một phần nhỏ của nghiên cứu mà có thể hưởng lợi từ FAIR. Điều này đã được khẳng định trong quá trình các cuộc phỏng vấn với các chuyên gia, những người đã chỉ ra rằng trong khi FAIR làm nản lòng việc giả mạo, nó cũng cải thiện chất lượng nghiên cứu trên diện rộng bằng việc cải thiện khả năng tái tạo lại.

3.5 Chỉ số #5: Cấp vốn hai lần

Chi phí cấp vốn hai lần vì nghiên cứu không FAIR là 25 triệu € mỗi năm. Bảng bên dưới trình bày dữ liệu đã được sử dụng để ước tính chỉ số này.

Đầu vào về đúp bản nghiên cứu	***Đơn vị***	***2017***
Tổng tỷ lệ phần trăm các cặp có sự trùng lặp đáng ngờ	%	0,03%
Đóng góp trung bình của EU vào các dự án được H2020 cấp vốn	EUR (€)	1.783.788€
Số lượng trợ cấp nghiên cứu công ước tính ở EU 28	#	189.014
Tổng số lượng trùng lặp đáng ngờ trong các giả thuyết về trợ cấp ở EU28	#	50
Tổng vốn cấp được phân bổ cho các cặp đáng ngờ	EUR (€)	89.185.346€
Kích cỡ trung bình của lần đầu trao trợ cấp của cặp đí so với lần thứ hai	#	1,9
Tổng giá trị vốn cấp cho các trợ cấp đúp bản	**EUR (€)**	**30.753.568€**
Giảm đúp bản nhờ có FAIR	%	80,00%
Chi phí cấp vốn hai lần có thể tránh được		***24.602.854€***

Bảng 6: Tính toán chỉ số #5

3.6 Chỉ số #6: Liên ngành

Tác động tiềm tàng lên liên ngành bị bỏ sót vì không có dữ liệu nghiên cứu FAIR có thể không được ước tính một cách tin cậy. Được thừa nhận rằng tác động của FAIR lên liên ngành, trên thực tế, không làm gia tăng tổng lượng nghiên cứu được thực hiện, mà thay vào đó cải thiện các khía cạnh đặc thù của bản thân nghiên cứu, như chất lượng kết quả đầu ra tốt hơn, cộng tác và khả năng sử dụng lại tốt hơn, .v.v. Trên cơ sở này, chúng tôi đã xác định vài yếu tố đã cho phép chúng tôi tin tưởng tác động hữu hình của FAIR lên liên ngành, mà còn cả sự đóng góp lợi ích đan xen cho chi phí của việc không có dữ liệu nghiên cứu FAIR:

- Liên ngành dựa một phần vào khả năng tái tạo lại và đòi hỏi sự minh bạch về các công cụ, phương pháp và dữ liệu được sử dụng. Điều này dẫn tới sự tin cậy ngày

một gia tăng của các phát hiện nằm bên trong các xuất bản phẩm khoa học. Nhờ có dữ liệu nghiên cứu FAIR, chất lượng nghiên cứu tổng thể được cải thiện.

- Đối với đại đa số các nhà nghiên cứu được khảo sát[82], việc thiếu truy cập tới dữ liệu và chất lượng dữ liệu yếu kém đang hạn chế liên ngành và cản trở chất lượng nghiên cứu tốt. Việc giới thiệu dữ liệu nghiên cứu FAIR có thể dẫn tới lợi ích đan xen lớn hơn.
- FAIR có thể cho phép các nhà nghiên cứu truy cập dữ liệu khác nhau từ các ngành khác, bằng cách đó trao cho họ cơ hội giành được những thấu hiểu mới và vì thế tạo thuận lợi cho việc chia sẻ kiến thức.

Để tóm tắt, liên ngành nhiều hơn thông qua các nguyên tắc FAIR có thể đưa các ngành khoa học lại gần hơn và làm gia tăng tỷ lệ sử dụng lại dữ liệu hiện hành[83], cho phép khai phá các khả năng mới vượt ra khỏi cách thức truyền thống tiến hành nghiên cứu, điều có thể làm lợi cho cộng đồng khoa học và các lĩnh vực khác được kết nối tới nó.

3.7 Chỉ số #7: Tăng trưởng kinh tế tiềm năng

Tăng trưởng kinh tế tiềm năng bị bỏ sót vì không có dữ liệu nghiên cứu FAIR có thể cũng không được ước tính một cách tin cậy. Tuy nhiên, chúng tôi đã xác định một số yếu tố cho phép chúng tôi tin tưởng các tác động của FAIR lan tỏa về kinh tế bị bỏ sót có thể tạo nên một thành phần lớn nhất về chi phí của việc không có dữ liệu nghiên cứu FAIR:

- FAIR và tính mở có tác động tích cực trực tiếp lên số lượng các trích dẫn. Nếu tất cả mọi điều ngang bằng như nhau, thì nghiên cứu FAIR sẽ được sử dụng lại thường xuyên hơn và vì thế có nhiều giá trị hơn.
- Các tài nguyên theo truy cập mở tạo nên nguồn thông tin được sử dụng nhiều thứ hai sau các bộ sưu tập từ cơ sở của chính mình, theo một khảo sát[84]. Bằng việc gia tăng khả năng tiếp cận và khả năng sử dụng lại dữ liệu nghiên cứu, FAIR

82 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3126798/table/pone-0021101-t008/ https://www.elsevier.com/about/open-science/research-data/open-data-report (Cox & Williamson, 2014)

83 (Womack, 2015)

84 http://www.sr.ithaka.org/wp-content/uploads/2016/06/SR_Report_UK_Survey_Academics_2015_06152016.pdf

cũng có thể làm gia tăng giá trị xã hội của nghiên cứu vì nghiên cứu mở được sử dụng lại thường xuyên hơn.

- Trong thiểu số các nhà nghiên cứu hiện đang làm cho dữ liệu sẵn sàng ở dạng điện tử cho những người khác, đa số làm như vậy vì họ được yêu cầu phải làm như vậy, theo một nghiên cứu[85]. Việc áp dụng các nguyên tắc FAIR ở các cơ sở hoặc ở mức của các nhà cấp vốn có thể vì thế cải thiện rất nhiều cho tính sẵn sàng của dữ liệu nghiên cứu, bằng cách đó tạo ra giá trị.
- Ngoài ra, bằng chứng thực nghiệm[86] cho thấy rằng phần các xuất bản phẩm theo truy cập mở đã ngày một gia tăng rồi kể từ những năm 1990. Trong trường hợp không có chính sách hay ưu đãi ràng buộc, điều này gợi ý mạnh mẽ rằng có những lợi ích để làm cho nghiên cứu truy cập được nhiều hơn.

Cuối cùng, vì nền kinh tế dữ liệu châu Âu được ước tính có quy mô gần bằng với chi phí nghiên cứu của châu Âu (cả hai đều vào khoảng 300 tỷ €[87]), nên có thể rút ra một phép so sánh với dữ liệu mở. Những lợi ích kinh tế có thể đối với dữ liệu mở đã được ước tính[88] giữa 11,7 tỷ € và 22,1 tỷ € mỗi năm ở châu Âu đến năm 2020 và vì thế có thể được kỳ vọng về những lợi ích kinh tế của FAIR nằm trong khoảng tương tự.

85 http://www.ijdc.net/index.php/ijdc/article/view/10.1.210/393
86 https://ec.europa.eu/research/openscience/index.cfm?pg=access&section=monitor#viz1489066430689
87 (European Commission, 2017) and (Eurostat, 2017)
88 (European Commission, 2017)

4. KẾT LUẬN

Việc diễn giải tổng chi phí của việc không có dữ liệu nghiên cứu FAIR dưới dạng một giá trị đơn lẻ sẽ bỏ qua nhiều lợi ích không định lượng được của FAIR. Dù vậy, ở mức 10,2 tỷ € mỗi năm ở châu Âu, chi phí đo lường được này của việc không có dữ liệu nghiên cứu FAIR tạo ra một trường hợp áp đảo ủng hộ cho việc triển khai các nguyên tắc FAIR.

Để hiểu rõ điều này, chi phí nghiên cứu ở châu Âu lên tới 302,9 tỷ € vào năm 2016. Trong khi chi phí thực tối thiểu của việc không có FAIR có thể thấy chỉ 3% của tất cả chi tiêu nghiên cứu, 10,2 tỷ € mỗi năm là 78% ngân sách của Horizon 2020 mỗi năm và là gần 400%, của những gì Hội đồng Nghiên cứu châu Âu và các hạ tầng nghiên cứu châu Âu nhận được cộng lại.

Ngoài ra, các số liệu về nền kinh tế dữ liệu mở gợi ý rằng tác động lên đổi mới sáng tạo của FAIR có thể bổ sung thêm 16 tỷ € khác vào chi phí tối thiểu chúng tôi đã ước tính.

Trong khi nghiên cứu này không tính tới chi phí triển khai FAIR, nếu chúng tôi giả thiết rằng các chi phí bổ sung được phân bổ cho quản lý dữ liệu chiếm tới 2,5% tất cả chi tiêu nghiên cứu, thì điều này sẽ để lại số dư dương ~ 2,6 tỷ € mỗi năm từ việc triển khai các nguyên tắc FAIR. Ngoài ra, không phải tất cả các chi phí cho việc triển khai các nguyên tắc FAIR sẽ được tái diễn. Một khi đã có được cơ sở hạ tầng thích hợp, người ta có thể kỳ vọng lợi ích ròng từ các nguyên tắc FAIR sẽ tăng lên.

Nghiên cứu của chúng tôi trình bày các kết quả cho nền kinh tế nghiên cứu của Liên minh châu Âu như một tổng thể, tuy nhiên chi phí của việc không có FAIR biến động mạnh từ ngành này qua ngành khác. Trong một vài ngành tăng cường dữ liệu như gen học hay tinh thể học, (vài trong số) các nguyên tắc FAIR đã được triển khai rồi mà không cần phân tích chi phí - lợi ích được định lượng.

Kết quả là, một số khoản khấu trừ có thể được rút ra từ công việc đã hoàn thành:

- FAIR là một phần của phong trào lớn hơn đang làm thay đổi cách thức khoa học được thực hiện, với sự nổi lên của quản trị dữ liệu và xung lượng đang gia tăng có lợi cho tính mở. Tương tự, là cơ bản rằng các hạ tầng và chính sách cần thiết được triển khai để hưởng lợi đầy đủ từ các nguyên tắc FAIR và tối đa hóa giá trị của dữ liệu nghiên cứu.

- Về khía cạnh chi phí của việc không có nghiên cứu FAIR, thời gian bỏ ra và chi phí lưu trữ là các trình điều khiển chi phí đo đếm được đáng kể nhất. Nói cách khác, FAIR có thể có tác động đáng kể lên thời gian chúng ta bỏ ra điều khiển dữ liệu và cách thức chúng ta lưu trữ dữ liệu.
- Do đó, chúng tôi tin tưởng rằng các nguyên tắc FAIR sẽ đóng góp lớn cho hệ sinh thái khoa học và đổi mới sáng tạo ở châu Âu, nhưng thiếu liên kết dữ liệu FAIR và đổi mới sáng tạo làm cản trở định lượng tác động của FAIR lên đổi mới sáng tạo.

Chi phí của việc không có nghiên cứu FAIR đã được tính toán từ các chỉ số định lượng: thời gian bỏ ra, chi phí lưu trữ, chi phí cấp phép, rút lại nghiên cứu và cấp vốn hai lần. Các chỉ số đó không thể bao trùm tất cả những lợi ích của FAIR lên đổi mới sáng tạo và tăng trưởng kinh tế. Ngoài ra, những giả thiết rất bảo thủ đã được sử dụng và vài hạn chế áp dụng cho các chỉ số đó, ví dụ:

- Với thời gian bỏ ra, chúng tôi đã không tính tới thời gian bị lãng phí vì không có FAIR có thể bị/được phát minh lại trong nghiên cứu, điều có thể dẫn tới hoàn vốn đầu tư nhất định.
- Với chi phí lưu trữ, các tính toán của chúng tôi chỉ bao trùm nghiên cứu hàn lâm, vì thiếu dữ liệu cho khu vực tư nhân.
- Với các chi phí cấp phép, các tính toán của chúng tôi chỉ bao trùm khung chi phí đối với tổ chức nghiên cứu công, vì thiếu dữ liệu cho khu vực tư nhân.
- Với rút lại nghiên cứu, chúng tôi loại bỏ một số lượng nghiên cứu không xác định không bị rút lại nhưng cũng có chất lượng kém và không thể tái tạo lại ở một mức độ nào đó do dữ liệu không là FAIR.

Chúng tôi cũng có thể không cân nhắc đầy đủ tác động của việc không có FAIR lên khả đọc được và khả năng sử dụng của máy. Lượng dữ liệu được sản xuất đang gia tăng hàng ngày và trong vài năm từ nay trở đi các nhà nghiên cứu sẽ không có khả năng làm việc và xử lý các lượng đó thủ công bằng tay. Nếu dữ liệu là FAIR, máy có thể hỗ trợ cho các nhà nghiên cứu trong việc tìm kiếm và phân tích dữ liệu nghiên cứu có thể tác động tích cực tới hệ sinh thái khoa học, về các khía cạnh tăng tốc thời gian, độ chính xác, lượng dữ liệu được phân tích, và tốc độ mà còn về những hiểu biết mới được rút ra.

Vì thế, chúng tôi tin tưởng rằng chi phí thực của việc không có dữ liệu nghiên cứu FAIR là lớn hơn nhiều so với được ước tính 10,2 tỷ € mỗi năm. Để ước tính nó đầy đủ, dữ liệu bổ sung cần phải được thu thập, đặc biệt cho nghiên cứu trong các tổ chức tư nhân.

Như là nhà sản xuất hàng đầu thế giới về lượng dữ liệu nghiên cứu, châu Âu đã đặt nghiên cứu vào cốt lõi của chiến lược phát triển của nó. Bằng việc mở khóa giá trị gia tăng từ dữ liệu nghiên cứu, việc áp dụng các nguyên tắc FAIR không chỉ là câu hỏi về hiệu quả. Việc không làm như vậy cũng gây ra một chi phí rõ ràng là chuyển hướng các nguồn lực từ bước đột phá khoa học tiếp theo.

TÀI LIỆU THAM KHẢO

Addis, M. (2015). Estimation of research data volumes per researcher in UK HEI. doi:10.6084/m9.figshare.1577539

American Journal Experts. (n.d.). Peer Review: How We Found 15 Million Hours of Lost Time. Retrieved from American Journal Experts: https://www.aje.com/en/arc/peer-review-process-15-million-hours-lost-time/

Baker, M. (2016). 1500 scientists lift the lid on reproducibility. Retrieved from https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Beagrie, N., & Houghton, J. (2014). The Value and Impact of Data Sharing and Curation. Retrieved from http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf

Beagrie, N., & Houghton, J. (2016). The Value and Impact of the European Bioinformatics Institute. Retrieved from https://beagrie.com/static/resource/EBI-impact-report.pdf

Biology, A. I. (2015). ASCB Member Survey on Reproducibility. Retrieved from http://www.ascb.org/wp-content/uploads/2015/11/final-survey-results-without-Q11.pdf

Björkdahl, J. (2009). Technology cross-fertilization and the business model: The case of integrating ICTs in mechanical engineering products. Research Policy, 1468-1477. Retrieved from https://www.researchgate.net/publication/46489026_Technology_cross-fertilization_and_the_business_model_The_case_of_integrating_ICTs_in_mechanical_engineering_products

Bonino da Silva Santos, L. O., Wilkinson, M., Kuzniar, A., Kaliyaperumal, R., Thompson, M., Dumontier, M., & Burger, K. (2016). AIR Data Points Supporting Big Data Interoperability. Retrieved from https://www.researchgate.net/publication/309468587_FAIR_Data_Points_Supporting_Big_Data_Interoperability

Charles Beagrie Limited. (2011). User Guide for Keeping Research Data Safe - Assessing Costs/Benefits of Research Data Management, Preservation and Re-use. Retrieved from https://beagrie.com/static/resource/KeepingResearchDataSafe_UserGuide_v2.pdf

Cheol Shin, J., Arimoto, A., Cummings, W. K., & Teichler, U. (2014). Teaching and Research in Contemporary Higher Education. Springer. Doi:10.1007/978-007-6830-7

Council of the European Union. (2016, May 27). OUTCOME OF PROCEEDINGS: The transition towards an Open Science system – Council conclusions. Retrieved from http://data.consilium.europa.eu/doc/document/ST-9526-2016-INIT/en/pdf

Court, S. (2012, October). An analysis of student:staff ratios and academics' use of time, and potential links with student satisfaction. Retrieved from https://www.ucu.org.uk/media/5566/An-analysis-of-studentstaff-ratios-and-academics-use-of-time-and-potential-links-with-student-satisfaction-Dec-12/pdf/ucu_ssranalysis_dec12.pdf

Cox, A., & Williamson, L. (2014). The 2014 DAF Survey at the University of Sheffield. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/10.1.210/393

Data Citation Synthesis Group. (2014, February). Joint Declaration of Data Citation Principles. (M. M., Editor, & F. 11, Producer) doi:10.25490/a97f-egyk

Data Life Cycle | DataONE. (n.d.). Retrieved January 2018, from DataOne | Data Observation Network for Earth: https://www.dataone.org/data-life-cycle

Davis, P. (2012). How Much of the Literature Goes Uncited? Retrieved from The Scholarly kitchen: https://scholarlykitchen.sspnet.org/2012/12/20/how-much-of-the-literature-goes-uncited/

Directorate-General for Research & Innovation (European Commission). (2017). Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Retrieved from

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Directorate-General for Research and Innovation (European Commission). (2017). The economic rationale for public R&I funding and its impact. Publications Office of the European Commission. Retrieved from https://publications.europa.eu/en/publication-detail/-/publication/0635b07f-07bb-11e7-8a35-01aa75ed71a1/language-en

Dunning, A., de Smaele, M., & Böhmer, J. (2017). Are the FAIR Data Principles fair? 4TU.ResearchData. Retrieved from https://zenodo.org/record/321423#.Wku301WnE3F

European Commission. (2017). Final results of the European Data Market study measuring the size and trends of the EU data economy. Retrieved from https://ec.europa.eu/digital-single-market/en/news/final-results-european-data-market-study-measuring-size-and-trends-eu-data-economy

European Commission. (2017). Review of the Directive on the re-use of public sector information (Directive 2013/37/EU). Retrieved from https://ec.europa.eu/info/law/better-regulation/initiatives/ares-2017-4540429_en

Eurostat. (2017). R & D expenditure. Retrieved from Eurostat: http://ec.europa.eu/eurostat/statistics-explained/index.php/R_%26_D_expenditure

Fang, F. C., Steen, G. R., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. Proceedings of the National Academy of Sciences of the United States of America, 17028-17033. doi:https://doi.org/10.1073/pnas.1212247109

Garner, H., McIver, L., & Waitzkin, M. (2013). Same work, twice the money? Nature, 493, 599-601. Retrieved from https://www.healthra.org/download-resource/?resource-url=/wp-content/uploads/2013/11/Garner_Nature_1_2013.pdf

Horbach, S., & Halffman, W. (2017). The extent and causes of academic text recycling or 'self-plagiarism'. Elsevier, 11. doi:https://doi.org/10.1016/j.respol.2017.09.004

Houghton, J., & Gruen, N. (2014). Open Research Data - Report to the Australian National Data Service (ANDS). Retrieved from http://apo.org.au/system/files/53613/apo-nid53613-72236.pdf

Johnson, R., Parsons, T., Chiarelli, A., & Kaye, J. (2016). Jisc Research Data Assessment Support - Finding of the 2016 data assessment framework (DAF) surveys. doi:10.5281/zenodo.177856

Johnson, R., Parsons, T., Chiarelli, A., & Kaye, J. (2016). JISC Research Data Assessment Support - Findings of the 2016 data assessment framework (DAF) surveys. doi:10.5281/zenodo.177856

Kejser, U. B., Nielsen, A. B., & Thirifays, A. (2011). Cost Model for Digital Preservation: Cost of Digital Migration. The International Journal of Digital Curation. Retrieved from http://www.ijdc.net/index.php/ijdc/article/viewFile/177/246

Lahti, L. (2016). Scientific journal subscription costs in Finland 2010-2015: a preliminary analysis. RopenGov. Retrieved from http://ropengov.github.io/r/2016/06/10/FOI/

Lawson, S., Meghreblian, B., & Brook, M. (2015). Journal subscription costs - FOIs to UK universities. Retrieved from figshare: https://figshare.com/articles/Journal_subscription_costs_FOIs_to_UK_universities/1186832

Masic, I. (2012). Plagiarism in scientific publishing. Acta Informatica Medica, 208-213. doi:10.5455/aim.2012.20.208-213

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L., & Wilkinson, M. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use, vol. 37, no. 1, pp. 49-56.

Nur Farah Naadia Mohd Fauzi, P., Abd Rashid, K., Ahmad Sharkawi, A., Farihah Hasan, S., & Aripin, S. (2016). A survey on required time allocated vs. actual time spent by

academic in pursuit of key performance indicators (KPIs) at departme,t of quantity surveying, KAED, IIUM. 42-48. Retrieved from http://jsrad.org/wp-content/2016/Issue%204,%202016/7jj.pdf

Palaiologk, A. S., Economides, A. A., Tjalsma, H. D., & Sesink, L. B. (2012). An activity-based costing model for long-term preservation and dissemination of digital research data: the case of DANS. International Journal on Digital Libraries. doi:10.1007/s00799-012-0092-1

Parsons, T., Grimshaw, S., & Williamson, L. (2013, 02 06). Research Data Management Survey. Retrieved from http://eprints.nottingham.ac.uk/1893/1/ADMIRe_Survey_Results_and_Analysis_2013.pdf

Peter, Larry, Raphael, Gary, & Beth. (2015). FAIR and RDA DFT: sharing the same key messages. Retrieved from https://b2share.eudat.eu/api/files/e2d84fea-b4e2-41d1-8d4e-6ce2d315b2b9/comparison-fair-dft-v2.pdf

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. Retrieved from https://peerj.com/articles/175/?utm_content=buffer9059d&utm_source=buffer&utm_medium=twitter&utm_campaign=Buffer

Royal Veterinary College, University of London. (n.d.). Costing Research Data Management. Retrieved from https://www.rvc.ac.uk/research/about/research-data-management/before-a-project/costing-research-data-management: https://www.rvc.ac.uk/research/about/research-data-management/before-a-project/costing-research-data-management

Stehouwer, H., & Wittenburg, P. (2014). RDA Europe: Data Practices Analysis. Research Data Alliance Europe. Retrieved from https://b2share.eudat.eu/api/files/0312c522-968c-43e2-8127-9772d68ba084/iCORDI-D2%205-final-submit.pdf

Tenopir, C., Allard, S., Douglass, K., Umur Aydinoglu, A., Wu, L., Read, E., . . . Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. PLoS ONE. doi:https://doi.org/10.1371/journal.pone.0021101

Van Noorden, R. (2013). Open access: The true cost of science publishing. Nature. Retrieved from https://www.nature.com/news/open-access-the-true-cost-of-science-publishing-1.12676

Van Tuyl, S., & Michalek, G. (2015). Assessing Research Data Management Practices of Faculty at Carnegie Mellon University. Doi:10.7710/2162-3309.1258

Wager, E., & Williams, P. (2011). Why and how do journals retract articles? An analysis of Medline retractions 1988-2008. Journal of medical ethics, 9. doi:10.1136/jme.2010.040964

Ware, M., & Mabe, M. (2012). The stm report - An overview of scientific and scholarly journal publishing. Retrieved from http://www.stm-assoc.org/2012_12_11_STM_Report_2012.pdf

Wilkison, M. D., Dumontier, M., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. doi:10.1038/sdata.2016.18

Wittenburg, P. (2017). Costs of FAIR Compliance and not being FAIR compliant. doi:10.23728/b2share.e184bd1ff12d45269de80c3f3e443eb7

Wittenburg, P., Ritz, R., & Berg-Cross, G. (2015). RDA Data Foundation and Terminology - DFT: Results RFC. Retrieved from https://b2share.eudat.eu/api/files/266e2ea2-6200-4b9d-9d3b-ed6d03ff93a5/DFT%20Core%20Terms-and%20model-v1-6.pdf

Womack, P. R. (2015). Research Data in Core Journals in Biology, Chemistry, Mathematics, and Physics. doi:10.1371/journal.pone.0143460

Zahedi, Z., Haustein, S., & Bowman, T. D. (2014). Exploring data quality and retrieval strategies for Mendeley reader counts. Retrieved from http://www.asis.org/SIG/SIGMET/data/uploads/sigmet2014/zahedi.pdf

Ziker, J. N. (2013, October 13). Time Allocation Workload Knowledge Study, Phase 1 Report. ResearchGate. Retrieved from https://www.researchgate.net/publication/308761975_Time_Allocation_Workload_Knowledge_Study_Phase_1_Report

Phụ lục I - CÁC NGUYÊN TẮC FAIR

Bảng bên dưới mô tả các tiêu chí khác nhau cần thiết để gắn với các nguyên tắc FAIR. Cột bên trái liệt kê 15 khía cạnh tương ứng với FAIR, trong khi cột bên phải trình bày các thước đo để đánh giá tuân thủ FAIR của một tài nguyên kỹ thuật số[89]. Từ '(siêu) dữ liệu' được sử dụng trong các trường hợp nơi mà các Nguyên tắc đó cần được áp dụng cho cả siêu dữ liệu và dữ liệu.

Bảng 7: Các nguyên tắc FAIR và các khía cạnh đối với một tập hợp (siêu) dữ liệu

Hướng dẫn các nguyên tắc FAIR	**Các thước đo FAIR được đề xuất[90]**
Để tìm thấy được (Findable):	
F1. (Siêu) dữ liệu được chỉ định một mã thường trực bên ngoài và độc nhất toàn cầu	*Tính độc nhất của mã nhận diện* Có một URL liên kết tới một hệ thống mã nhận diện được đăng ký, xác định duy nhất tài nguyên số đó. Vài ví dụ không vét cạn là RN, IRI, DOI, Handle, trustyURI, LSID, .v.v. *Tính thường trực của mã nhận diện* Có một URL liên kết tới một chính sách có thể quản lý những thay đổi trong hệ thống mã nhận diện
F2. Dữ liệu có mô tả siêu dữ liệu phong phú (được R1 bên dưới xác định)	*Khả năng máy đọc được siêu dữ liệu* Các thuộc tính để tối ưu hóa phát hiện chúng. Các thuộc tính có cấu trúc và có siêu dữ liệu phong phú như tiêu đề, người tạo lập, ngày xuất bản (các) từ khóa, độ phủ về thời gian và không gian, .v.v. Lý tưởng, một URL nên liên kết tới tài liệu chứa siêu dữ liệu máy đọc được đối với tài nguyên số.
F3. (siêu) dữ liệu có đăng ký và được đánh chỉ mục trong một tài nguyên tìm kiếm được.	*Mã nhận diện tài nguyên trong siêu dữ liệu* Siêu dữ liệu phải chứa mã nhận diện cho tài nguyên kỹ thuật số mà nó mô tả một cách rõ ràng, do đó phải cung cấp URL của siêu dữ liệu và IRI (mã nhận diện tài nguyên được quốc tế hóa) của tài nguyên kỹ thuật số mà nó mô tả.
F4. Siêu dữ liệu chỉ định mã nhận diện dữ liệu	*Được đánh chỉ mục trong một tài nguyên tìm kiếm được* Mã nhận diện thường trực của tài nguyên và một hoặc nhiều URL đưa ra các kết quả tìm kiếm của các máy tìm kiếm khác nhau phải được cung cấp.

89 https://www.go-fair.org/fair-principles/
https://zenodo.org/record/321423#.WhV7oFWnE3E
https://www.nature.com/articles/sdata201618
https://www.force11.org/group/fairgroup/fairprinciples
https://via.hypothes.is/https:/content.iospress.com/articles/information-services-and-use/isu824#ref019

90 https://github.com/FAIRMetrics/Metrics/blob/master/ALL.pdf

Để truy cập được (Accessible):	
A1. (siêu) dữ liệu mà mã nhận diện của chúng truy xuất được bằng việc sử dụng một giao thức truyền thông được tiêu chuẩn hóa	*Giao thức truy cập* Phải cung cấp một URL mô tả giao thức dù đó là giao thức mở và miễn phí, giao thức đóng, giao thức có tiền bản quyền. *Ủy quyền truy cập*
A1.1 giao thức đó là mở, miễn phí, và triển khai vạn năng	Trong trường hợp các hạn chế, giao thức theo đó nội dung có thể truy cập được phải được chỉ định đầy đủ (nghĩa là, liệu ủy quyền có cần thiết và mô tả quy trình để truy cập được tới nội dung).
A1.2 Giao thức đó cho phép một thủ tục ủy quyền và xác thực khi cần	
A2. (Siêu) dữ liệu đó là truy cập được, kể cả khi dữ liệu đó không sẵn sàng	*Tuổi thọ siêu dữ liệu* Một siêu dữ liệu trỏ tới một kế hoạch tuổi thọ của siêu dữ liệu chính thức phải được cung cấp, vì siêu dữ liệu phải duy trì là phát hiện được, kể cả khi không có dữ liệu đó.
Để tương hợp được (Interoperable):	
I1. (Siêu) dữ liệu sử dụng ngôn ngữ chính thức, truy cập được, được chia sẻ, và áp dụng được rộng rãi để trình bày kiến thức	*Sử dụng ngôn ngữ biểu diễn kiến thức* Cần thiết sử dụng các ngôn ngữ có khả năng trình bày các khái niệm theo cách máy hiểu được. Một URL liên kết tới đặc tả của một ngôn ngữ như vậy phải được cung cấp.
I2. (Siêu) dữ liệu sử dụng từ vựng tuân theo các nguyên tắc FAIR	*Sử dụng các từ vựng theo FAIR* Bản thân các giá trị siêu dữ liệu và quan hệ đủ điều kiện phải là FAIR, ví dụ, các khái niệm từ các từ vựng mở, được cộng đồng chấp nhận, được xuất bản trong một định dạng trao đổi kiến thức thích hợp. Sau đó, phải cung cấp một IRI đại diện cho các từ vựng được sử dụng cho (siêu) dữ liệu.
I3. (Siêu) dữ liệu gồm các tham chiếu đủ điều kiện tới các (siêu) dữ liệu khác	*Sử dụng các tham chiếu đủ điều kiện* Mối quan hệ trong (siêu) dữ liệu, và giữa dữ liệu địa phương và bên thứ ba, có ý nghĩa về ngữ nghĩa rõ ràng và 'hữu ích'.
Để sử dụng lại được (Reusable):	
R1. (Siêu) dữ liệu có nhiều thuộc tính chính xác và phù hợp	Các khía cạnh của siêu dữ liệu giúp người ta đánh giá một tập hợp dữ liệu sử dụng lại được như thế nào. Các tác giả không nên cố gắng xác định các cách sử dụng xuôi dòng có thể, mà nên cung cấp càng nhiều thuộc tính càng tốt, ngoài các thuộc tính cần thiết cho việc sử dụng xuôi dòng dự kiến của họ.
R1.1 (Siêu) dữ liệu được phát hành với một giấy	*Giấy phép sử dụng truy cập được* (Siêu) dữ liệu được phát hành với một giấy phép sử dụng dữ liệu rõ ràng và

phép sử dụng dữ liệu rõ ràng và truy cập được	truy cập được. Một IRI của giấy phép đó (nghĩa là, URL của nó) cho giấy phép dữ liệu và cho giấy phép siêu dữ liệu đó phải được cung cấp. Có khả năng máy đọc được là một điểm cộng và có thể bằng cách tham chiếu tới một trong các giấy phép tại địa chỉ: http://purl.org/NET/rdflicense
R1.2 (Siêu) dữ liệu được liên kết tới xuất xứ chi tiết	*Xuất xứ được chi tiết hóa* Thông tin xuất xứ như ai/cái gì/khi nào đã sản xuất dữ liệu và vì sao/làm thế nào dữ liệu đã được sản xuất phải có liên kết tới dữ liệu đó. Hai URL phải được cung cấp. Một trỏ tới các từ vựng được sử dụng để mô tả xuất xứ trích dẫn, cái kia trỏ tới một trong các từ vựng (có khả năng đặc thù lĩnh vực) được sử dụng để mô tả xuất xứ theo ngữ cảnh.
R1.3 (Siêu) dữ liệu đáp ứng các tiêu chuẩn cộng đồng phù hợp lĩnh vực	*Đáp ứng các tiêu chuẩn cộng đồng* Cung cấp một chứng thực, từ một cơ quan được thừa nhận, nói rằng tài nguyên này tuân thủ với các tiêu chuẩn cộng đồng.

Liên hệ với Liên minh châu Âu

GẶP TRỰC TIẾP

Khắp Liên minh châu Âu có hàng trăm Trung tâm Thông tin Trực tiếp của châu Âu. Bạn có thể thấy địa chỉ của trung tâm gần bạn nhất tại: http://europa.eu/contact

TRÊN ĐIỆN THOẠI HOẶC QUA THƯ ĐIỆN TỬ

Europe Direct là một dịch vụ trả lời các câu hỏi của bạn về Liên minh châu Âu. Bạn có thể liên hệ dịch vụ này:

- bằng điện thoại miễn phí: 00 800 6 7 8 9 10 11 (một vài nhà vận hành có thể lấy tiền đối với các cuộc gọi đó),
- bằng số tiêu chuẩn sau đây: +32 22999696 hoặc
- bằng cách qua thư điện tử: http://europa.eu/contact

Tìm kiếm thông tin về Liên minh châu Âu

TRÊN TRỰC TUYẾN

Thông tin về Liên minh châu Âu trong tất cả các ngôn ngữ chính thức của Liên minh châu Âu sẵn có trên website Europa tại địa chỉ: http://europa.eu

CÁC XUẤT BẢN PHẨM CỦA LIÊN MINH CHÂU ÂU

Bạn có thể tải về hoặc đặt hàng các xuất bản phẩm của Liên minh châu Âu miễn phí hoặc mất tiền từ các cửa hàng sách tại: http://bookshop.europa.eu. Nhiều bản sao các xuất bản phẩm miễn phí có thể có được bằng cách liên hệ với Europe Direct hoặc trung tâm thông tin địa phương của bạn (xem http://europa.eu/contact)

CÁC TÀI LIỆU LUẬT CỦA LIÊN MINH CHÂU ÂU HOẶC CÓ LIÊN QUAN

Để truy cập tới thông tin pháp lý từ Liên minh châu Âu, bao gồm tất cả luật của Liên minh châu Âu từ 1951 trong tất cả các ngôn ngữ chính thức, hãy tới EUR-Lex tại địa chỉ: http://eur-lex.europa.eu

DỮ LIỆU MỞ TỪ LIÊN MINH CHÂU ÂU

Cổng Dữ liệu Mở của Liên minh châu Âu (http://data.europa.eu/euodp/en/data) cung cấp truy cập tới các tập hợp dữ liệu từ Liên minh châu Âu. Dữ liệu có thể được tải về và sử dụng lại miễn phí, cả vì các mục đích thương mại và phi thương mại.

Dữ liệu nghiên cứu FAIR xoay quanh cách thức để tạo lập, lưu trữ và xuất bản dữ liệu nghiên cứu theo cách thức chúng là tìm thấy được, truy cập được, tương hợp được và sử dụng lại được. Để là FAIR, dữ liệu nghiên cứu được xuất bản phải đáp ứng các tiêu chí nhất định được các nguyên tắc FAIR mô tả. FAIR bắt nguồn từ các thực hành quản lý dữ liệu một cách chắp vá hiện hành ở Liên minh châu Âu, điều còn chưa là tối ưu. Vài sáng kiến địa phương, cũng như toàn cầu, đang dịch chuyển hướng tới một hạ tầng hỗ trợ các nguyên tắc FAIR để có được nhiều nhất dữ liệu nghiên cứu. Báo cáo này nhằm ước tính chi phí của việc không có dữ liệu nghiên cứu FAIR đối với nền kinh tế của Liên minh châu Âu dựa vào một loạt các chỉ số đo lường được, chúng đã được xác định dựa vào các nghiên cứu và các cuộc phỏng vấn hiện hành với các chuyên gia về chủ đề này.

Các nghiên cứu và báo cáo

Văn phòng Xuất bản của Liên minh châu Âu