氏　　名　南山 泰之

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2366 号

学位授与の日付　2022 年 9 月 28 日

学位授与の要件　複合科学研究科 情報学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Analysis and Formalization of Data Curation Process for Open Science

論 文 審 査 委 員　主　査　武田 英明
情報学専攻 教授
北本 朝展
情報学専攻 教授
岡田 仁志
情報学専攻 准教授
大向 一輝
東京大学 大学院人文社会系研究科 准教授
高久 雅生
筑波大学 図書館情報メディア系 准教授

（様式３）

博士論文の要旨

氏　　名　南山　泰之

論文題目　Analysis and Formalization of Data Curation Process for Open Science

With the expansion of international collaborative research and the trend of open science, expectations have been raised for research data sharing and reuse across fields. For research data reuse, data curation is essential to make the data interpretable and reusable. In some leading fields, field-specific tasks and procedures have been developed to implement systematic data curation. However, even in those leading fields, the research data reuse is often closed within the field. For interdisciplinary reuse of research data, published research data must be interpretable by researchers from different fields. The problem here is the difference in data curation, which depends on the field. This variation in the data curation by field reduces the interpretability of research data. Without a method to overcome this challenge, open science will not be realized. This study will analyze the practices of data curation in each field for interpreting the data curation process in a cross-disciplinary perspective. Furthermore, we will provide technical support for interpreting the data curation process with a formalized method.

In Chapter 2, we provide an overview of the research background and related areas. First, we review the current status of research data sharing and reuse with its social positioning. Then we discuss the significant studies from the data curation field. Finally, we show our approach to address the issues.

In Chapter 3, we formalize the knowledge of the data curation process to provide a cross-disciplinary protocol across different fields for computational processing. The granularity of tasks and procedures that constitute data curation building blocks is not formalized, so it is impossible to identify common tasks and procedures across different fields. Here we propose an approach using ontology theory and techniques to interpret data curation tasks and procedures across fields. Our proposed ontology will allow data re-users to interpret the tasks and procedures performed with a cross-disciplinary protocol. This means eliminating potential risks such as improperly misusing metadata for reuse and research data reuse. Our approach promotes re-organizing the data curation process in an interpretable across different fields. This study helps the data re-users to interpret the procedural aspects of the data curation.

In Chapter 4, we deal with the processes in the Data Evaluation category formalized in Chapter 3 through practical implementations. To interpret research data from a field to be independently, the data and their documentation must be reviewed from a cross-disciplinary perspective and revised as necessary. On the other hand, many data

repositories either do not have a policy for evaluating research data and encouraging improvement, or the evaluation process is done only from the perspective of a specific field. This chapter focuses on publishing mechanisms for data papers that divert the journal peer-review and clarifies the mechanism from a process perspective. Furthermore, we conduct a technical investigation for implementing the mechanism using the reference model for data publishing. This study is a practical examination of the relationship and boundaries between the existing quality assurance process and the peer-review process in the Data Evaluation category.

In Chapter 5, we deal with the processes in the Appraisal category formalized in Chapter 3, focusing on the processes related to setting conditions of use. These processes have different variations due to the differences in legal restrictions and disciplinary norms by jurisdiction. Also, the conditions of use granted by data providers are more diverse. As a result, it is difficult for data re-users to interpret the results of the process accurately. Research data reuse based on ambiguous interpretation of processing results often leads to unintended reuse for the data providers. This chapter investigates the actual processing status related to setting conditions of use in different fields and clarifies the correspondence of information tied to each formalized process. Furthermore, we conduct a technical investigation for stepwise interpretation of these processing results. This study is positioned as an effort to complement chapter 4 practices each other and support the building of a cross-disciplinary infrastructure from a rule perspective.

In Chapter 6, we discuss the results of this study and prospects. Through this thesis, we provided our framework for viewing data curation activities as processes from a cross-disciplinary perspective. This framework allows the data curation process to be interpreted in a decoupled way of the original research context. Data re-users will be able to formally assess the increased interpretability by verifying that the processes included in the framework have been properly executed. Furthermore, we have demonstrated some practical implementations through our problem-solving approach as a stepwise formalization to the level of interpretation. We believe that as we move forward with these formalized efforts, we can improve the interpretability of research data and thereby contribute to the realization of open science.

Form8・Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏　　名　　南山　泰之

Title
論文題目　　Analysis and Formalization of Data Curation Process for Open Science

本学位論文は、「Analysis and Formalization of Data Curation Process for Open Science」と題し、全６章から構成されている。

第１章「Introduction」では、本論文の背景から始め、研究の動機を述べた上で、研究構成について述べている。近年、オープンサイエンスの潮流の中で研究データの公開が学術活動の重要な要素として認識され実践されるようになった一方、そのためのプロセスであるデータキュレーションは分野ごとの取り組みに留まっている。オープンサイエンスが目指す分野を超えたデータ利用にはこのデータキュレーションを分野を超えて理解できるようにする必要がある。このため、本論文は、研究データのデータキュレーションのプロセスを分析して形式化するとともに、その一部のプロセスについては実践的なプロセスの設計を行うことを目的としている。

第２章「Research background and related area」では関連研究を述べている。各分野での研究データのキュレーションの実際を述べるとともに、データキュレーションプロセスについて学術的な試みについて紹介している。その上、本論文の目的に対する取り組み方法を述べている。

第３章「Formalizing the knowledge of data curation process across different fields」ではデータキュレーションプロセスの形式化に関する研究を述べている。まず、データキュレーションの実際を把握するため、国内８研究機関でのデータキュレーションの実際をインタビュー調査を行った。その上で、人工知能の一分野である知識工学で使われているオントロジーによる形式化を行なっている。オントロジー構築の手順に基づいて、調査した情報を構造化して、データキュレーションプロセスオントロジーを構築した。このオントロジーは 184 個のクラスと 85 個の関係、33 個の属性からなっている。このオントロジーに基づき､調査を行った研究機関のデータキュレーションプロセスを記述することができ、機関間でのプロセスの差異を明示化できるなど、オントロジーの効用を示した。

第４章「Practical interpretation of the cross-disciplinary data evaluation processes using the reference model for data publishing」では、データキュレーションプロセスの１プロセスであるデータ・パブリッシングについて、具体的なデータジャーナルでの査読プロセスを取り上げ、その設計について述べている。データ・パブリッシングの一つの方法としてデータ論文という方式が行われている。本研究では、データ論文の出版において論文の改訂とデータの改訂が並列的に行うプロセスを設計し、実際のデータジャーナルの査読プロセスに採用されたことを述べている。

第５章「Reframing the cross-disciplinary appraisal processes by analyzing conditions of use of research data」では、データキュレーションプロセスの１プロセス

であるアプレイザル（査定）について、実践的な設計を行なっている。このプロセスに関して、実務者へのインタビュー及び関係者へのアンケート調査を行い、リスク・マネージメント・プロセスとライツ・マネージメント・プロセスに大別されること、またそれぞれにおいて必要となる条件を明らかにした。この結果はガイドラインとしてまとめられ、外部機関により公開されている。

第６章「Conclusion」では、これまでの研究成果をまとめるとともに今後の展望について述べている。

公開発表会では博士論文の章立てに従って発表が行われ、その後に行われた論文審査会及び口述試験では、審査員からの質疑に対して適切に回答がなされた。

質疑応答後に審査委員会を開催し、審査委員で議論を行った。審査委員会では、出願者の博士研究が学際的な研究としての独自性・独創性と今日の学術活動に対する貢献があることが評価された。

以上を要するに本学位論文は、データキュレーションというオープンサイエンスに必須な活動を形式化及び実践で探求した論文であり、その独自の取り組みそのものが学際的な研究活動として高く評価できることに加え、その成果は今日の科学の健全な発展に寄与する重要な貢献があると言える。また、本学位論文の成果は、学術雑誌論文２件として発表され、社会的な評価も得ている。以上の理由により、審査委員会は、本学位論文が学位の授与に値すると判断した。