

研究データ管理オンライン講座の開発と受講者特性の分析

古川雅子^{†1} 尾城孝一^{†1} 山地一禎^{†1}

概要：研究データ管理は、研究プロセスの透明性を高め不正を防止するとともに、オープンサイエンスにおける研究データ共有を支える重要な基礎となる。しかし、日本において、研究データ管理のスキルを持つ人材は必ずしも十分とは言えないのが現状である。本稿では、このような人材育成を目指して開発した研究データ管理オンライン講座について述べる。また、事前/事後アンケート及び学習ログの分析により、受講者の特性や行動の傾向を明らかにする。具体的には、各週の映像のアクセス数を開講期間の4週分並べ、この4次元データを個人ごとの特徴量とし、k平均法によりクラスターに分割した。この結果、映像の視聴回数が多いクラスターは合格者が多く、またこのコースのメインのターゲットである図書系の受講者は、このクラスターの割合が高いことが分かった。小テストのアクセス数についても同様の分析を行い、図書系の受講者は、熱心に映像の視聴や小テストの試行を行うことや成績が良い受講者の割合が高いことが明らかになった。また、受講者の9割が、講座が有用であったと答え、研究データ管理に関する具体的な知識を得ることができた点などを評価していることが明らかになった。

キーワード： オープンサイエンス、ラーニングアナリティクス、MOOC、研究データ管理

Development and Analysis of RDM Training Online Course

MASAKO FURUKAWA^{†1} KOICHI OJIRO^{†1}
KAZUTSUNA YAMAJI^{†1}

Abstract: Research data management (RDM) is the basic skill for promoting research data sharing in open science, as well as for enhancing transparency in research process. However, its training materials and environment are still under development in Japan. In this paper, we aimed to develop and provide the RDM training online course. In order to clarify the learning behavior of participants, pre and post questionnaire and the learning log were analyzed. Number of video view for 4 weeks was taken as a feature of each participant, and categorized as clusters by the k-means method. As a result, the clusters with a large number of video view showed high completion rate, and majority of this cluster was librarian. Similar analysis was carried out for the number of quiz access. These lead us to the conclusion that enthusiastically accessed to the training materials by librarians result high completion rate and their high expectation to RDM. In addition, 90% of the participants evaluated that the lecture was useful, and satisfied to the fact that specific knowledge on RDM was obtained than before.

Keywords: Open Science, Learning Analytics, MOOC, Research Data Management

1. はじめに

2013年のG8科学大臣会合における研究データのオープン化に関する共同声明を皮切りとして、国内でもオープンサイエンスに関する議論が活発化している[1]。論文だけではなく、データについても広く容易に再利用可能とすることで、研究の加速化と不正防止という両側面でのメリットが生まれる。この双方を実現するために不可欠となるのが、研究を遂行する段階からの適切な研究データ管理 (Research Data Management: RDM) である。研究データ管理とは、ある研究プロジェクトにおいて使用された、あるいは生成されたデータの組織化、構造化、保存、共有、公開、再利用に関する一連の作業を指す。

2018年6月29日には、内閣府の国際的動向を踏まえたオープンサイエンスの推進に関する検討会から、国立研究開発法人におけるデータポリシー策定のためのガイドライ

ンが公開された[2]。こうしたデータポリシーの策定は、研究機関だけではなく、大学にも展開されることが予想される。先行するイギリスのエジンバラ大学では、研究データを保存することだけではなく、それを共有して、公開して、再利用できるようなポリシーを大学として作成している。研究前、研究中、研究後、日常的な教育支援の取り組みという形で、研究者をサポートするサービスを大学が組織として提供している[3]。

このようなサービスを実現するためには、具体的に何をどのようにサポートすべきか知らなくてはならないが、海外では、研究データに関して様々なオンライン教材が公開されている。エジンバラ大学 MANTRA は、研究プロジェクトの一環としてデジタルデータを管理する人を対象とした無料のオンラインコースを提供している[4]。FOSTERは、オープンサイエンスの詳細を知るために必要なeラーニングコースが集められている[5]。Figshareは、研究データを

^{†1} 国立情報学研究所
National Institute of Informatics

公開するためのレポジトリであり、研究データ管理に関する教材も公開されている[6].

日本における教材開発に関しては、オープンアクセスリポジトリ推進協会(JPCOAR)の研究データタスクフォースが、海外の動向も調査しながら研究データ管理の基礎を学ぶ教材の開発を行ってきた[7]. 初期段階の教材が開発された段階にあり、今後は、研究データ管理に関する基礎知識を必要とする支援員や研究者に普及させていく必要がある。比較的大規模な展開が必要となることに加え、教材を改善していくための適切な環境を整備していくことが、研究データ管理に関する教育的側面における今後の課題となっている。特に、すでに開発した教材を実際に関係者に提供し、受講者の特性を考慮しながら、教材の改善やサポート体制を検討材料としていくことが、今後の展開を見据える上でも重要な意味をもつ。

本研究では、研究データ管理に関するこうした環境整備の一環として、オンライン講座の開発と提供を実施する。オンライン講座には、比較的大規模な講座を提供するための広報やシステム環境が既に用意されている、大規模公開オンライン講座(MOOC)を活用する。一般社団法人日本オープンオンライン教育推進協議会(JMOOC)[8]が提供するプラットフォームでは、学習ログなどの情報を提供するサービスがあることから[9]、ここで得られた受講者の行動履歴やアンケート結果を分析することで、今後の研究データ管理における教育を実践していく上での基礎的な情報を獲得する。

2. 研究データ管理オンラインコースの開発

この教材のねらいは、学習者が研究データ管理に関する基礎的な知識を習得することと、研究データ管理サービス構築の足掛かりを得ることである。このコースのメインのターゲットは、研究データ管理支援を担うと期待されている大学・研究機関の図書系職員である。

JPCOARが開発したRDMトレーニングツールは、全7章の音声付きのeラーニング教材として作成され、各章は、スライドと解説と確認テストで構成される。研究データのライフサイクル(生成, 加工, 分析, 保存, 公開, 再利用)をサポートできる教材として作成され、各章の構成は、以下ようになる。

第1章: 導入編であり、RDMが必要とされる背景や、研究データおよびRDMの定義について解説している。

第2章: DMP(Data Management Plan)の定義や動向について述べた上で、実際のDMPの策定方法を解説する。あわせてDMPを支援するツールについても紹介している。

第3章: 研究データの保存と共有をテーマとし、長期保存にあたってのセキュリティ上の留意事項や、共有・再利用のためのデータリポジトリの活用などについて解説して

いる。

第4章: 研究データの組織化、文書化、メタデータ作成をテーマとし、主要なメタデータスキーマを紹介するなど、データを再利用に供するためのシステムティックな管理方法を解説している。

第5章: 法・倫理的問題をテーマとし、著作権に関するライセンスや、研究不正に関するポリシーについて解説している。

第6章: 研究者が研究を進めるにあたってのポリシーをテーマとしており、研究公正に関するポリシーの事例と、そこでの研究データ取扱いについて解説している。

第7章: RDMサービスの設計をテーマとし、RDMを支援する人材の確保や、研究データ保存・公開の基盤システムについて解説している。

これらは、2017年6月6日付でJPCOARの公式ウェブサイト上でCC-BYで公開された。

RDMトレーニングツールの公開後、研究データタスクフォースでは、ツールのさらなる有効活用を目指し、国立情報学研究所と共同でMOOCプラットフォームのgaccoを利用して、RDMトレーニングツールを活用した講座を開講した。このMOOCコース「オープンサイエンス時代の研究データ管理」は、2017年11月15日から2018年1月15日まで開講された。

図1は、その受講画面を示している。RDMトレーニングツールでは7章で構成されていた教材をMOOCのために4週間のコースに再編成し、研究データタスクフォースの協力を得て映像を補完した。MOOCの1週間分にはRDMトレーニングツールの2章分が含まれるが、4章の分量が多かったため、4章の内容は、第3週と第4週に分けられた。また、JPCOARのRDMトレーニングツールでは、音声を付与したスライドのみであったため、JMOOC講座の一般的な映像のように講師が登場する部分を挿入し、JMOOCの学習者になるべく違和感を与えないよう配慮をした。具体的には、教材音声のナレーターと教材作成者が担当した章について簡単な紹介をするシーンを挿入し、その後にはナレーターのスライド音声が続くような構成とした。

このMOOCコースの各週は、4~5本の映像、ダウンロード教材、内容確認テスト(10問の選択問題)で構成される。このほか、MOOCコースの機能として、任意で回答する開始前アンケートと実施後アンケートおよび、ディスカッションボードが用意された。今回開講したコースでは、各週の確認テストの合計点が7割に達した場合に修了証を発行した。

開発したMOOCコースの第1週では、研究データ管理の重要性が増している背景や研究データ管理の意義について学ぶという内容になっている。第2週では、研究データの保存と共有、文書化について学習する。第3週では、メタデータ・法・倫理的問題について学ぶ。第4週では、研

研究データに関するポリシーと、研究者が研究データを適切に管理するために、サービスを組織としてどのように設計していけば良いのかについて学ぶ。



図 1 研究データ管理オンラインコース
Figure 1 RDM training online course.

3. 研究データ管理オンラインコースの分析

3.1 データ分析プラットフォーム

学習解析プラットフォームの概要を図 2 に示す。学生が LMS 上で、コース中の動画再生や小テストといったモジュールを使用するたびに、学習ログの新しい入力在学习管理システムに追加される。学習ログには、閲覧時間、ログイン時間の総数、オンラインディスカッションの総数、レポートの採点結果などが含まれる。

開発したシステムは、xAPI という学習ログの標準に基づいて構築されており、ダッシュボード上では、統計解析等で広く利用されている R を用いて詳細な分析等を行うことができる。また、作成した分析コードを共有化する機能もあり、NII がこれまで行ってきた機関レポジトリの機能と合わせて、教育コンテンツの蓄積、分析、共有のためのプラットフォームとして利用することができる。

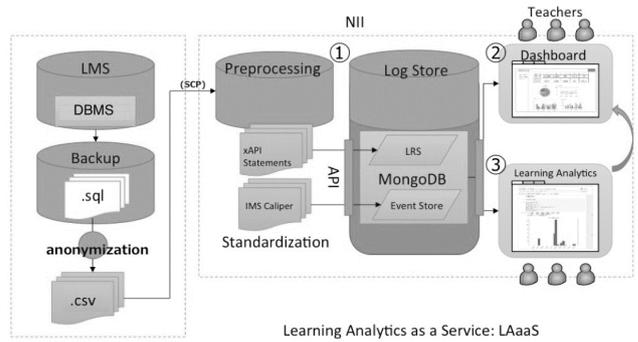


図 2 データ分析プラットフォーム
Figure 2 Data analysis platform.

3.2 研究データ管理オンラインコースの分析

開発した研究データ管理オンラインコースの受講者数は、2,305 名だった。直近 1 年の講座平均受講者数は 4,145 名である。平均と比べると約半分であるが、gacco では一般的な教養内容が多い中、専門性の高い内容であることから当初は 800 人程度の受講者数を見込んでおり、予想を上回ったと言える。また、修了率は 25% であり、gacco の平均修了率は 15%、MOOC の世界的なレベルでの修了率も 10% 前後という中で高いと言える (表 1)。

表 1 受講者数と修了率

Table 1 Number of attendants and completion rate.

	受講者数	ディスカッション スレッド数	修了率
オープンサイエンス時代の研究データ管理	2,305	13	25%
gacco 講座平均 (昨年平均)	4,145	73	15%

得点分布を見ると、登録のみで受講しないことによる 0 点を除くと、合計 100 点で修了した受講者が最も多かった。得点を取っている受講者は 70% が修了のラインであるのにも関わらず 100 点を目指しているという特徴があった。これはすべての週の確認テストで同じ傾向があった (図 3)。

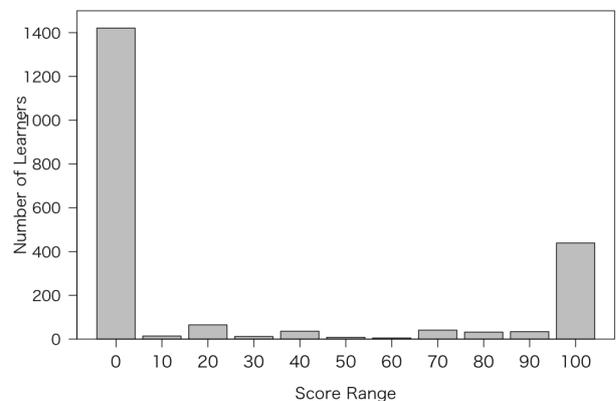


図 3 得点の分布
Figure 3 Distribution of points.

開始アンケートについては、回答は必須ではないものの、770名の回答を得た。回答者の6割程度が男性であり、年齢層はほぼ全体の構成と同じだった。職種は、フルタイムと回答した者が616名であった。勤務先は、大学研究機関が50%を占めた。その中でも特に多かったのは図書系と回答した者で、回答者全体の30%を占めていた(図4)。

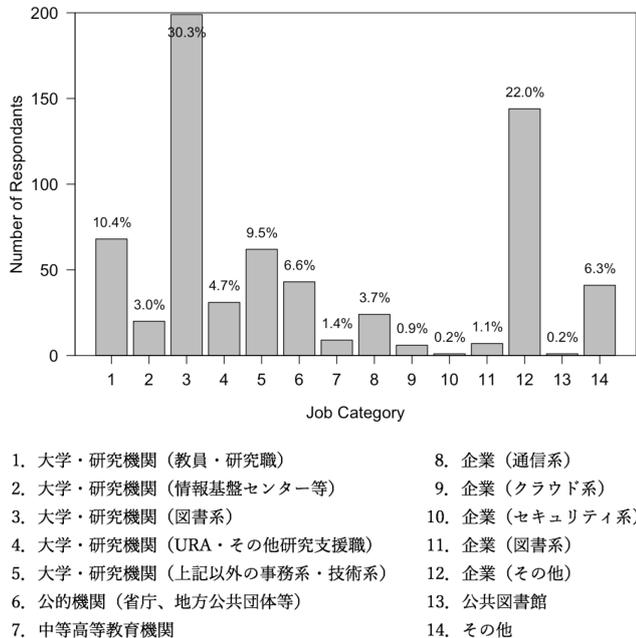


図4 回答者の職種

Figure 4 Job category of respondents.

次に教材のアクセス数について分析を行う。開発したコースは、映像をメインとして構成されていることから、受講者によって映像の視聴傾向に違いがあるかを分析の対象とした。受講者のクラスタリングには、広く用いられる手法の1つであるk平均法を利用した。具体的には、各週の映像のアクセス数を、開講期間の4週分並べ、この4次元データを個人ごとの特徴量とした。そして、この4次元データをk平均法により、クラスタに分割した。

k平均法における適切なクラスタ数を決めるための方法の1つとしてエルボー法があるが、これは、クラスタ数を小さくしながら、データとクラスタ重心の二乗誤差が急に大きくなる直前を適切なクラスタ数とするものである。しかし、今回分析したデータでは、映像の再生エラーなどによりアクセス数が極端に増加する場合があったため、エルボー法を用いると、データ数がごく小さい、外れ値が属するいくつかのクラスタと、ほとんどのデータが属する1つのクラスタに分離された。このため、クラスタ数を大きくしながら、大きなクラスタが3から4に分かれるという条件で、クラスタ数を15と設定した。

図5にクラスタごとの合格、不合格数を示す。外れ値に対応するクラスタは、構成人数が少なくなっている。人数

が多いクラスタは、クラスタg, h, lである。そして、クラスタgは、3割程度が不合格であったのに対し、クラスタh, lは、合格する割合が高いことが分かる。

クラスタg, h, lの重心を見ると、最も多くの受講者が含まれるクラスタgの各週の映像のアクセス数は平均5.4回と、必ずしも多くはなかった。一方、h, lのクラスタを見ると、各週の映像のアクセス数は平均20.5回、47.1回と、映像の視聴回数が多いクラスタであった。

職種ごとの各クラスタの人数を図6に示す。このコースのメインのターゲットである図書系(3)を見ると、映像の視聴回数が多いクラスタであるh, lに属する人数が多く、熱心に映像の視聴を行なっていることが分かる。

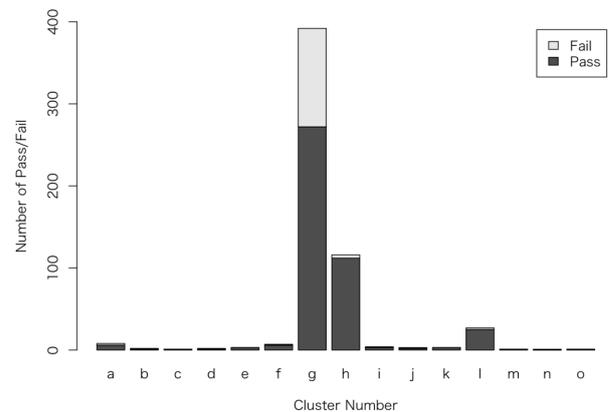


図5 クラスタごとの合格/不合格数(映像視聴)

Figure 5 Pass / fail number of each cluster (Movie).

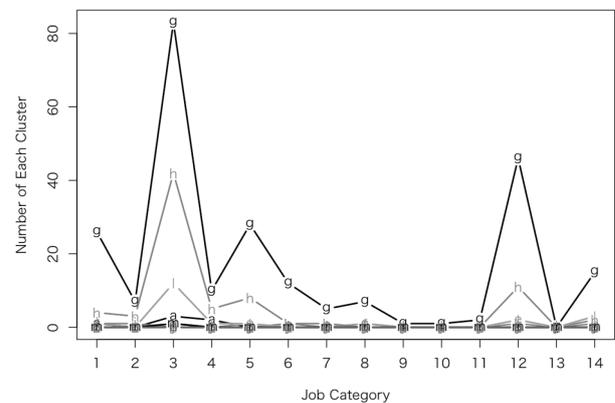


図6 職種ごとの各クラスタの人数(映像視聴)

Figure 6 Number of people in each cluster for each job category (Movie).

小テストの試行回数についても、映像のアクセス数と同様の分析を行った。図7は、各週の小テストの試行回数を、開講期間の4週分並べ、この4次元データをk平均法により、クラスタに分割した結果である。映像のアクセス数と同様に、外れ値となるデータがあったことから、kの値を変えながら、外れ値以外の大きなクラスタが3から4に分

かれるという条件で、クラスタ数を5と設定した。クラスタ y, z を見ると、このクラスタは、合格する割合が高いクラスタであることが分かる。y, z のクラスタの重心を見ると、小テストの試行回数が多いクラスタであった。

また、職種ごとの各クラスタの人数を図8に示す。このコースの主なターゲットである図書系(3)を見ると、特に、小テストの試行回数が多いクラスタである y に属する人数が多く、小テストを熱心に試行している割合が高いことが分かる。

以上により、図書系の受講者は、熱心に映像の視聴や小テストの試行を行う成績が良い受講者の割合が多いことが分かる。

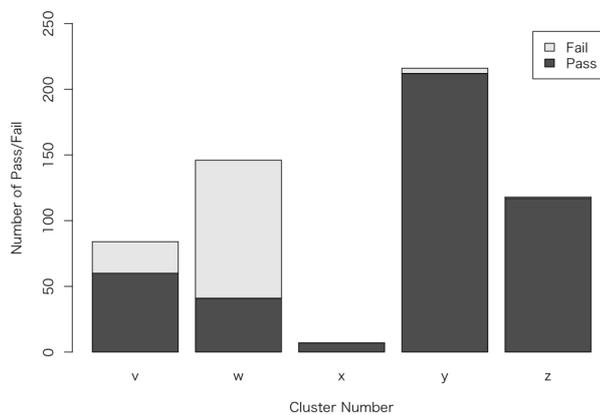


図7 クラスタごとの合格/不合格数(小テスト)
Figure 7 Pass / fail number of each cluster (Test).

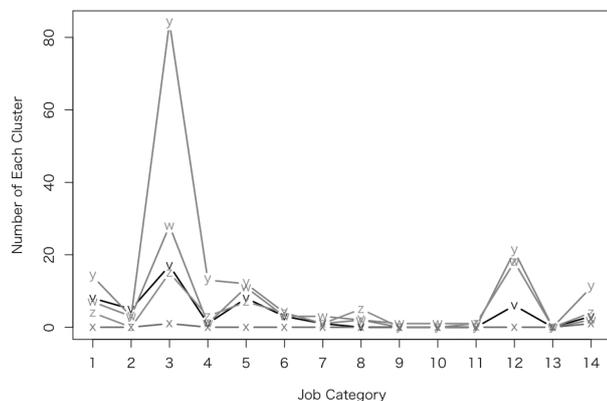


図8 職種ごとの各クラスタの人数(小テスト)
Figure 8 Number of people in each cluster for each job category (Test).

受講後アンケートに関しては、回答者数は345名だった。受講後アンケートは、第4週がアクセス可能になった時に回答可能になる。性別と年齢層、勤務先は、開始前アンケートとほぼ同じ構成だった。

「講座の内容は、あなた自身にとって有用でしたか」について質問したところ、「大変有用である」、「有用である」

の割合は9割程度であった(図9)。図書系である職種3とそれ以外を比較すると、図書系の受講者は、「大変有用である」と答えた割合が高かった。

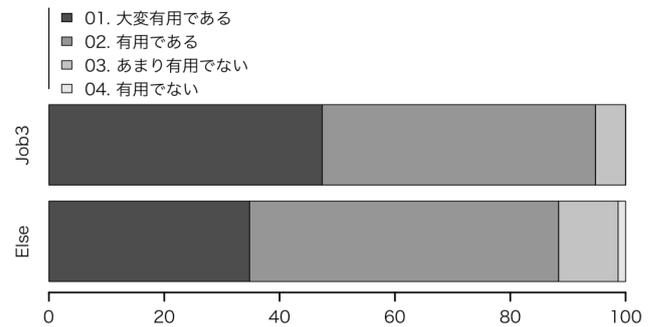


図9 有用であったか否か
Figure 9 Whether it was useful or not.

また、講座の内容は有用であったか否かについて、自由記述により、そのように回答した理由を聞いた。図書系でこの間に答えた人数は116名、それ以外で答えた人は233名であった。それぞれの内容の違いを見るために、それぞれの自由記述の内容を共起ネットワークにより分析した。共起ネットワークは、同時に出現する確率の高い単語同士を線で繋ぐことで、単語間の関係を可視化する手法であり、自由記述の分析等で利用される。

図10は、図書系の受講者の回答について共起ネットワークを作成したものである。この図と元の記述を見ると、例えば、「オープンサイエンスの基本的な考え方について学ぶことができた」、「図書館の現状について認識できた」といった記述がある反面、「図書館と他の部局との連携が課題」といった内容の記述があることが分かる。

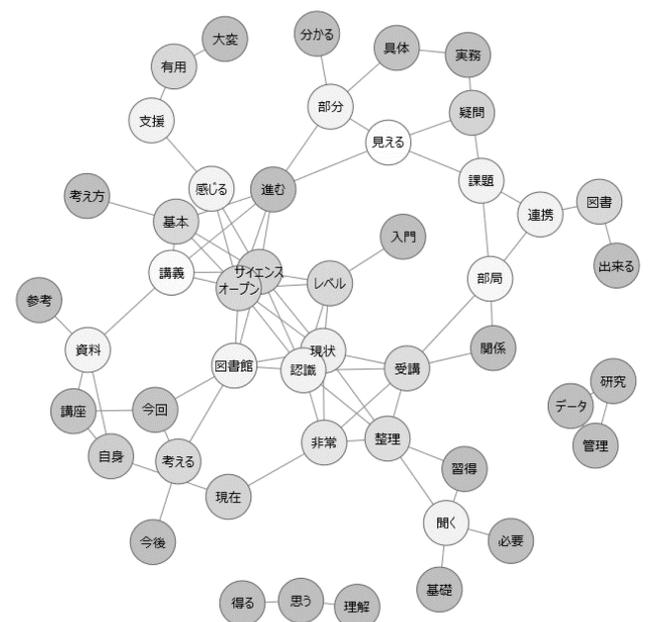


図10 共起ネットワーク(職種3)
Figure 10 Co-occurrence network (Job 3).

図 11 は、図書館以外の受講者の回答について共起ネットワークを作成したものである。この図と元の記述を見ると、例えば、「今回、勉強して RDM について分かった」、「仕事に役立つ」といった記述がある反面、「大学として整備することは困難」といった、現状の難しさについて言及する記述が見られた。

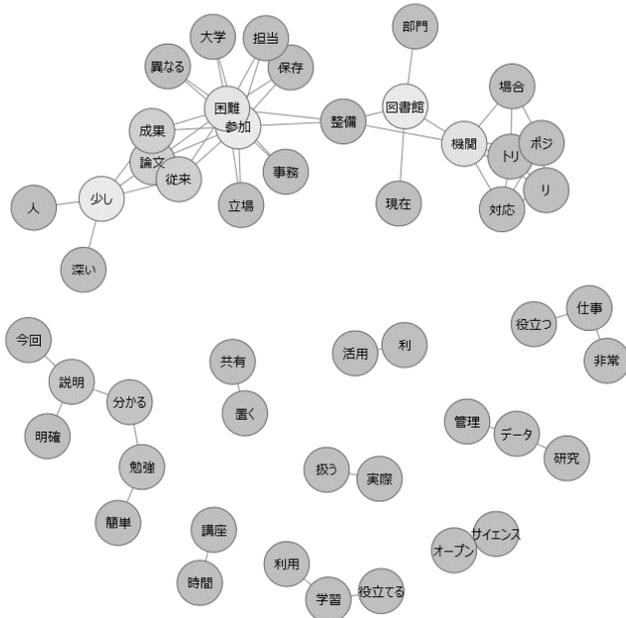


図 11 共起ネットワーク（それ以外）
Figure 11 Co-occurrence network (Else).

4. まとめ

本稿では、研究データ管理オンラインコースの開発について述べるとともに、事前/事後アンケート及び学習ログの分析を行った。分析の結果、受講者のうち半分程度が大学・研究機関関係者であり、特に図書館職員が 30%を占めること、図書館系の受講者は、熱心に映像の視聴や小テストの試行を行う、成績が良い受講者の割合が高いことが分かった。また、受講者の 9 割が、講座が有用であったと答え、RDM に関する具体的な知識を得ることができた点などを評価していることが明らかになった。

現在、研究データ管理サービスの設計と実践という仮題で、研究支援者、研究者を支援する立場の職員、基盤センター技術系スタッフなどの支援者向けに、研究プロセス(研究前、研究中、研究後)に沿ってどのようなサービスをしたらいいのか、そのデザインはどうしたらいいのか、実践するにはどうしたらいいのかなどを学べる教材を企画している。今後、今回の分析結果を見ながら、2018 年度内の開講を目標に準備を進めていく予定である。

参考文献

- [1] 船守美穂. 2017. オープンサイエンス推進に関わる学術機関の役割と課題. 情報知識学会誌, vol.27, no.4, pp.309-322. DOI: https://doi.org/10.2964/jsik_2017_034
- [2] 国立研究開発法人におけるデータポリシー策定のためのガイドライン (Published June 29, 2018 by Cabinet Office, Government of Japan) , <http://www8.cao.go.jp/cstp/stsonota/datapolicy/datapolicy.pdf>
- [3] Rice, R. and Haywood, J. 2011 Research Data Management Initiatives at University of Edinburgh. The International Journal of Digital Curation, Vol.6, No2, pp.232-244. DOI: <https://doi.org/10.2218/ijdc.v6i2.199>
- [4] MANTRA, <https://mantra.edina.ac.uk/>
- [5] Orth, A., Pontika, N., Ball, D. 2016. FOSTER's Open Science Training Tools and Best Practices. IOS Press. DOI: <https://doi.org/10.3233/978-1-61499-649-1-135>
- [6] Figshare, <https://knowledge.figshare.com/open-data/about-rdm>
- [7] 常川真央, 天野絵里子, 大園隼彦, 西園由依, 前田翔太, 松本侑子, 南山泰之, 三角太郎, 青木学聡, 尾城孝一, 山地一禎. 2017. 研究データ管理(RDM)トレーニングツールの構築と展開. 情報知識学会誌, vol.27, no.4, pp.362-365. DOI: https://doi.org/10.2964/jsik_2017_042
- [8] JMOOC, <https://www.jmooc.jp/>
- [9] Furukawa, M. and Yamaji, K. 2017. Adaptive Recommendation of Teaching Materials Based on Free Descriptions in MOOC Course. Proceedings of the 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pp.1011-1012: DOI: <https://doi.org/10.1109/IIAI-AAI.2017.176>