Linking Educational Resources on Data Science

José Luis Ambite,[1] Jonathan Gordon,[2]* Lily Fierro,[1] Gully Burns,[1] Joel Mathew[1]
[1] USC Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, California 90292, USA
[2] Department of Computer Science, Vassar College, 124 Raymond Ave, Poughkeepsie, New York 12604, USA
ambite@isi.edu, jgordon@vassar.edu, {lfierro, burns, joel}@isi.edu

Abstract

The availability of massive datasets in genetics, neuroimaging, mobile health, and other subfields of biology and medicine promises new insights but also poses significant challenges. To realize the potential of big data in biomedicine, the National Institutes of Health launched the Big Data to Knowledge (BD2K) initiative, funding several centers of excellence in biomedical data analysis and a Training Coordinating Center (TCC) tasked with facilitating online and in-person training of biomedical researchers in data science. A major initiative of the BD2K TCC is to automatically identify, describe, and organize data science training resources available on the Web and provide personalized training paths for users. In this paper, we describe the construction of ERuDIte, the Educational Resource Discovery Index for Data Science, and its release as linked data. ERuDIte contains over 11,000 training resources including courses, video tutorials, conference talks, and other materials. The metadata for these resources is described uniformly using Schema.org. We use machine learning techniques to tag each resource with concepts from the Data Science Education Ontology, which we developed to further describe resource content. Finally, we map references to people and organizations in learning resources to entities in DBpedia, DBLP, and ORCID, embedding our collection in the web of linked data. We hope that ERuDIte will provide a framework to foster open linked educational resources on the Web.

1 Introduction

We have embarked on an ambitious program to identify, describe, and organize web-based learning resources on data science. Our current collection contains over 11,000 resources, described under a uniform schema (using Schema.org terminology) and an openly available topic ontology (DSEO, cf. Sect. 3). To create this collection, known as the Educational Resource Discovery Index for Data Science (ERuDIte), we have developed and applied methods from machine learning, knowledge representation, information retrieval, and natural language processing. To make our collection available, reusable, and human- and machine-readable, we have released our collection of resources as linked data (Heath and Bizer 2011) with references to well-known entities in DBpedia (Auer et al. 2007), DBLP (Ley 2002), and ORCID (orcid.org). Here, we detail the methods applied to specific problems in building ERuDIte. First, we provide an overview of our project. In later sections, we describe our resource collection effort, the resource schema, the topic ontology, our linkage approach, and the linked data resource provided.

The National Institutes of Health (NIH) launched the Big Data to Knowledge (BD2K) initiative to fulfill the promise of biomedical "big data" (Ohno-Machado 2014). The NIH BD2K program funded 15 major centers to investigate how data science can benefit diverse fields of biomedical research including genetics, neuroimaging, mobile health, and precision medicine. Ensuring that the advances produced by these centers and other research efforts yield the expected benefits for human health requires a significant increase in the number of biomedical researchers trained in data science. To address this need, the NIH has funded the BD2K Training Coordinating Center (TCC).

Data science is a rapidly evolving, interdisciplinary field that draws from statistics, machine learning, high-performance computing, databases, and knowledge from specific scientific domains. Given the growing popularity of data science, many open learning resources have been published on the Web. However, these resources vary greatly in quality, topic coverage, difficulty, and presentation formats, which may be confusing and daunting for learners. To address these challenges, the BD2K TCC is developing a web portal, **BigDataU.org**, powered by ERuDIte, to provide a dynamic, personalized educational experience for biomedical researchers interested in learning about data science. To build ERuDIte, we are developing novel, automated methods to identify, collect, integrate, describe, and organize learning resources (Ambite et al. 2017). Automated methods are critical for scalability and to track new developments in this rapidly changing field. Essentially, we are using data science to build our index of data science resources.

2 Identifying/Collecting Learning Resources

For our purposes, we define a learning resource as any online material useful in learning about data science. Our core criteria for resource selection are relevance to data science, quality, and pedagogical value. Initially, we focused

*This work was primarily performed while JG was at USC/ISI.

Provider/Source	Types	Total	With Descriptions	With Transcripts	With Slides or Documents
BD2K	Video, Written	681	602	277	72
edX	Course, Video	89	88	69	53
Coursera	Course, Video	256	256	81	83
Udacity	Course, Video	17	17	17	0
Videolectures.net	Video	8,577	6,166	7,994	4,699
YouTube	Video	988	873	749	0
ELIXIR	Course, Written	237	48	0	0
Bioconductor	Course, Written	5	2	0	0
Cornell Virtual Workshop	Course, Written	38	19	0	0
OHBM	Video	78	6	0	51
NIH	Video	1	1	0	0
Bioinformatics.ca	Course, Video	86	63	0	0
Google Books	Written	267	213	0	0
Total		**11,320**	**8,354**	**9,187**	**4,958**

Table 1: Currently Indexed Learning Resources

on known high-quality sources, including massive open online courses (MOOCs), such as Coursera, Udacity, and EdX, scientific conference talks and seminars from aggregators such as videolectures.net, and materials generated by BD2K Centers. We extract the metadata for these resources, including titles, descriptions, instructors and their affiliations, and related materials such as syllabi, slides, and videos (cf. Sect. 3). Some sources provide such data through public APIs (e.g., coursera.org and udacity.com). However, most sources require web scraping, for which we developed a modular framework, using the popular Python packages BeautifulSoup and Dryscrape, to handle both static and dynamic, JavaScript-based webpages. To date, we have collected a total of 11,320 learning resources, which vary in granularity from individual videos to online courses that include multiple video lectures and associated materials. Table 1 describes the current sources, the number of learning resources per source, and the types of information extracted, such as resource descriptions, video transcripts, and supporting slides or other written materials.[1]

YouTube videos. Recently, we have used machine learning techniques to identify high-quality learning resources from large open collections such as YouTube. To find relevant YouTube videos, we search for terms related to data science, drawn from mainly the *Domain* dimension of the DSEO (Sect. 3), for example: "bioinformatics" or ("regression" AND ("data science" OR "machine learning")). We executed 98 such queries and collected metadata from the videos and playlists appearing in the first 20 pages of results for each query, yielding a dataset of 122,557 unique videos. We manually annotated 2,298 videos, sampled from across the different pages of results for different queries, yielding 1,217 high-quality videos and 1,081 low-quality ones. We trained a random forest classifier using the extracted metadata, including titles, descriptions, video transcripts, number of views, and ratings, which achieved a F_1 score of 0.82, which is comparable to human inter-annotator agreement.

Google Books. For pedagogical reasons, we initially focused on video materials, but have started to extend our collection to scientific books. We issued a set of 54 queries (similar to those for YouTube) on the open Google Books API, which yielded 19,612 books. We collected both the metadata for each book (title, authors, description, publisher, URL) and snippets of text from within the book that surrounded hits of the search terms. To remove off-topic books, we generated a 200-topic latent Dirichlet allocation (LDA) topic model using the MALLET toolkit (McCallum 2002) and manually examined the word distributions of each topic to determine whether it was relevant to data science. We then removed documents with an "irrelevant topic" in any of its top three topics. We manually assessed the documents generated and found the relevance acceptable. This procedure yielded 12,379 book records available for human curation.

3 Describing Learning Resources

Learning Resources Metadata Schema. We have designed a common schema to represent the metadata of the learning resources in ERuDIte. We first reviewed existing standards, such as Dublin Core, Learning Resource Metadata Initiative (LRMI), IEEE's Learning Object Metadata (LOM), eXchanging Course Related Information (XCRI), Metadata for Learning Opportunities (MLO), and Schema.org vocabularies. Later, based on our collaboration with the ELIXIR consortium[2] and participation in the W3C Schema Course Extension Group, we mapped our schema to the Schema.org vocabulary defined by BioSchemas.org, preserving custom properties only when critically needed. Schema.org has the support of major search engines, which facilitates the discovery and dissemination of resources indexed in ERuDIte.

The key classes of our standard are *CreativeWork* (for learning resources), *Person* (for instructors or material creators), and *Organization* (for affiliations and resource providers). Ontology definition files and a graphical visu-

[1] The YouTube and Google Books collections described below are under curatorial review and are not fully included in Table 1.

[2] ELIXIR is a large effort that seeks to provide a distributed infrastructure for life-science data across Europe. ELIXIR's Training e-Support System (TeSS) plays a role analogous to the BD2K TCC.

alization of our standard appear at https://bioint.github.io/erudite-training-resource-standard. Listing 1 shows sample JSON-LD markup for an educational resource.

```
{"@context": { "@vocab": "https://schema.org/",
               "bdu": "http://bigdatau.org/",
               "dseo": "http://bigdatau.org/dseo#" },
  "@id": "bdu:resource/12379054539352678334",
  "@type": "CreativeWork",
  "author": [ { "@type": ["Person"],
                "@id": "bdu:person/Brian_Caffo" },
              { "@type": ["Person"],
                "@id": "bdu:person/Jeff_Leek" },
              { "@type": ["Person"],
                "@id": "bdu:person/Roger_D._Peng" } ],
  "description": "Linear models, as their name...",
  "genre": [ "dseo:advanced" , "dseo:regression",
             "dseo:video", "dseo:written_documents" ],
  "provider": [ { "@type": ["Organization"],
                  "@id": "bdu:organization/Coursera"},
                 { "@type": ["Organization"],
                  "@id": "bdu:organization/
                          Johns_Hopkins_University"} ],
  "name": "Regression Models",
  "url": "https://www.coursera.org/learn/
          regression-models"}
```

Listing 1: Example JSON-LD for a learning resource

Data Science Educational Ontology. To design the DSEO, we combined top-down and bottom-up approaches. We first identified relevant concepts based on our knowledge of the data science domain and existing categorizations, such as those of videolectures.net, and organized them hierarchically along six dimensions. Then we applied two bottom-up, semi-automated methods to refine and extend the ontology. First, we extracted noun phrases (from parse trees constructed by the Stanford Parser (Chen and Manning 2014)) and common bigrams and trigrams from the text associated with the resources (titles, descriptions, syllabi, transcripts, slides, etc.). We manually reviewed 8,160 such extractions and selected a total of 861 candidate concepts. Second, we used non-negative matrix factorization (Shahnaz et al. 2006) to discover topics in our resources. We analyzed the most significant words associated with each topic to define a concept for each of the topics. Much of this analysis confirmed the concepts identified earlier, but it also yielded ten additional concepts. We defined the following criteria for a concept to be included in DSEO:

1. Is the concept relevant for at least five resources?
2. Does the concept capture an abstracted phrase or idea that cannot be easily found by an information retrieval search over resource text?
3. Would the concept help a user find a resource?
4. Does a clear definition for the concept exist?
5. Can the concept be automatically predicted/learned?

Using these criteria, we reduced the ontology to a total of 126 concepts, which we organized hierarchically along the following six dimensions, which address specific questions:

Data Science Process (7) What stage of the data science process will this resource help me with?

Domain (83) What is the topic/field of this resource?

Datatype (18) What types of data are addressed?

Programming Tool (14) What programming tool is used?

Resource Format (2) How is this resource presented?

Resource Depth (2) How advanced is this resource?

Visualizations of all concepts for the six dimensions are available at bigdatau.org/explore_erudite.

DSEO is formally a Simple Knowledge Organization System (SKOS) vocabulary, with the hierarchical relationships encoded by the *skos:broaderTransitive* property. DSEO is publicly available at bioint.github.io/DSEO and bioportal.bioontology.org/ontologies/DSEO. (For ease of exploration on BioPortal, we defined a version using *rdfs:subClassOf*). DSEO needs to adapt to innovations in data science; as we index more resources, we will add more concepts.

Dimension	Training/CV Set Size	Testing Set Size
Domain	7,904	1,885
Resource Depth	1,241	299
Resource Format	7,870	1,989
Data Science Process	1,725	447
Programming Tool	429	109
Datatype	1,866	466

Table 2: Assigning DSEO concepts to learning resources

Tagging Learning Resources with DSEO concepts. For scalability, we use machine learning to automatically assign candidate concepts from DSEO to learning resources. These proposed assignments are later curated by human experts before adding them to resources' descriptions. We trained several classifiers over a manually curated gold-standard set of DSEO concept assignments. Table 2 shows the total number of learning resources used for training, cross-validation, and testing for each DSEO dimension. We only include tags that have a minimum of five examples of support. We handle the hierarchy of DSEO by assigning all ancestor tags to a resource. Thus, if a resource is tagged with "machine learning," we also tag it with its ancestors: "artificial intelligence," "probability statistics," "computer science," and "mathematics." Our approach is based on text classification. For each resource, we combine the title, subtitle, description, syllabus, transcript, and text from slides or other written documents into a single document. The input features for machine learning consist of a bag-of-words TF–IDF vector representation of these documents. The most successful classifiers were one-vs-all logistic regression with L1 regularization and one-vs-all random forest. We used scikit-learn (Pedregosa et al. 2011) for training and cross-validating with a multi-label stratified five-fold approach.

Table 3 shows the performance of the best classifier for each dimension on the test set for concepts with more than 5 resources of support. Our performance measure is the F_1

score, which is the harmonic mean of precision (positive predictive value) and recall (sensitivity). We report a weighted average F_1 score, with the weights equal to the number of true positives of each tag in the test set. This performance is sufficient to send the DSEO concept assignment for human curation. We have developed a web-based curation interface, which is used internally in the project. However, we envision opening it to users of the web portal or crowdsourced workers, allowing us to re-train and validate our automated tagging algorithms at greater scale. As more resources are curated, we expect the classification performance to improve.

Dimension	Classifier Type	P	R	F_1
Domain	Logistic Regression	0.74	0.88	0.80
Resource Depth	Random Forest	0.66	0.91	0.76
Resource Format	Logistic Regression	1.00	1.00	1.00
Data Science Proc.	Logistic Regression	0.69	0.77	0.73
Programming Tool	Logistic Regression	0.80	0.71	0.74
Datatype	Logistic Regression	0.75	0.86	0.79

Table 3: Precision, recall, and F_1 scores for the best classifier for each DSEO dimension.

4 ERuDIte as Linked Data

To make our ERuDIte index more useful, we have embedded it into the web of linked data. We matched (textual) mentions of people and organizations to DBpedia and DBLP, using entity linkage techniques (Winkler 1999; Naumann and Herschel 2010). These sources are central in the world of linked data and provide good coverage of entities of interest. DBpedia covers most of the organizations (often universities or large companies with research departments) and also contains famous researchers. However, the standard of notability for inclusion on Wikipedia and, therefore, DBpedia means that there is low coverage for the instructors and authors of online learning resources for data science. Therefore, we also matched people to DBLP, a comprehensive database of published computer science research. Through these two sources we also get mappings from people to their ORCIDs, which enables further linkage to scientific literature. We essentially solve four entity linkage problems, BigDataU–BigDataU, BigDataU–DBpedia, BigDataU–DBLP, and DBpedia–DBLP, simultaneously.

Finding entity references in BigDataU resources. We first identify distinct mentions of persons and organizations in ERuDIte's learning resources. For *Person* instances, we consider their name and affiliation, e.g., ("Andrew Ng," "Stanford University"). For *Organization* instances, we consider their name and web address, e.g., ("University of Southern California," "https://www.usc.edu"). We could have considered additional information such as resource text (e.g., title, description, etc.) to be compared with a person's research areas (e.g., from their papers in DBLP), but simpler techniques worked adequately. In our current dataset, we identified 1,190 organization and 6,876 person references.

Finding candidate entities in DBpedia and DBLP. For each reference to a person or organization, we search existing entity linkage services for DBpedia and DBLP for candidate entities. In addition to the name as it occurs in the original reference, we also search for simplified versions that remove common elements such as academic titles, e.g., "Prof." or "Ph.D.", or organizational types, e.g., "Inc." or "GmbH." For organization names, we attempt to remove additional location information or subdivision names. When a reference includes both a spelled-out name and its acronym, e.g., "University of Southern California (USC)," we consider them both together and independently. To ensure high recall of relevant entities from DBpedia, we search it in three ways:

1. We query the DBpedia spotlight (Daiber et al. 2013) entity linking service. This service tends to match substrings of longer names, e.g., given "University of Southern California," it will annotate it only with the *Southern California* entity. We use a relevance threshold of 0.3 and a *Person* type restriction when querying for a reference to a person and no restriction for organization queries.
2. We query DBpedia's lookup service (github.com/dbpedia/lookup). We restrict searches on person references to subclasses of *Person* in DBpedia's ontology. For organizations, we find the application of types on DBpedia to be inconsistent (e.g., the Big Data to Knowledge project is labeled as a *Band*), therefore, we query both with and without the *Organisation* type restriction.
3. We construct a DBpedia URL based on the exact name and dereference it. Surprisingly, this identified some entities not found by the other methods.

When a DBpedia result redirects to another entity, we follow that redirect. We currently ignore disambiguation pages. We use type constraints in DBpedia to filter the results. We keep a list of good types (e.g., *University*, *Scientist*), and bad types (i.e., unlikely to be an instructor in data science; e.g. *SportsTeamMember*). We remove instances of bad types, unless they are also instances of a good type. As an additional practical restriction, we remove matches of people deceased before 1980 as they are unlikely to be the authors of online learning resources.

For DBLP, we similarly combine two methods of matching entities: (1) we construct and dereference DBLP author URLs to find exact matches, and (2) we use DBLP's author search API which provides fuzzy name matches. From these candidate entities, we retrieve their canonical name, known aliases (e.g., Wikipedia redirects), description, canonical homepage, associated URLs, and, for persons, known affiliations. We also retrieve ORCIDs directly from DBLP, and from DBpedia through mappings via Wikidata.

Linking external entities. If a BigDataU entity maps to multiple external entities (DBpedia, DBLP), we check if they can be consistently linked. This is trivially done for persons that have ORCID mappings. If both entities map to the same ORCID, the mappings are confirmed. If such additional mappings are not present, we compare the information about the external entities based on the Monge–Elkan (Monge and Elkan 1997) hybrid similarity measure, which computes a set-based similarity score over the sequence-based similarity scores computed by a secondary similarity measure, in our case Jaro–Winkler similarity (Winkler 1990). For persons, we compute the similarity of their names and affiliations (with a weighting of 0.7 and 0.3 respectively). We select the strongest match by computing the similarity for the Cartesian product of the names and aliases for the entities and for all affiliation and homepage strings, including the same simplifications we used when searching for the entities. If the resulting score is greater than 0.9, then the entities are linked.

Linking BigDataU references to external entities. Finally, we link each of the BigDataU references to a person or organization to the DBpedia and DBLP entities, using the Monge–Elkan similarity metric. For a person, we compute the similarity of the name and affiliations to those of the external entities that were found when searching for this reference. If the reference does not include an affiliation, the comparison is done only for the reference name and the entity's name and aliases. Otherwise, 70% of the score is from the best name similarity and 30% from the best affiliation similarity. For organizations, it is 70% name similarity and 30% URL similarity. If the best aggregate Monge–Elkan similarity score is greater than 0.7, the reference is linked to the entity. Otherwise, it is considered ambiguous and is left as an unlinked text string. The descriptions of learning resources, persons, and organizations are updated to link to the relevant external entities using *schema:sameAs* properties, and exported as JSON-LD documents.

Evaluation. For a random sample of 100 unique references to organizations (out of 1,190) and 200 unique references to persons (out of 6,876), the authors manually verified the automatically derived mappings to DBpedia and DBLP, with 3 annotators for each organization reference and 2–3 annotators per person reference. The average inter-annotator agreement and the agreement between the record linkage predictions and the annotators is shown in Table 4.

	Organization	Person	
	DBpedia	DBpedia	DBLP
System to Human	0.803	0.963	0.911
Human to Human	0.887	0.983	0.943

Table 4: Average pairwise agreement for linking entities.

To compute the accuracy of our automatic linking, we created a gold standard where at least $\frac{2}{3}$ or all of the human annotators agreed. Table 5 shows the accuracy of our automated linking for mappings at different levels of annotator agreement (e.g., more than 2 out of 3 annotators agreed on 191 of the 200 persons matches to DBLP, and all of the annotators agreed on 188 matches).

Releasing ERuDIte as Linked Data. In the spirit of open data sharing, we expose all the metadata for each of the learning resources in the ERuDIte collection as linked data in the JSON-LD format, under a Creative Commons Attribution-ShareAlike 4.0 International license.

We provide the ERuDIte linked data in two ways. First, every learning resource page on BigDataU.org includes a JSON-LD representation of the resource metadata (cf. Listing 1). This facilitates indexing by search engines and thus greater accessibility for users. Using the resource metadata, we provide faceted search on the TCC Web Portal (e.g., http://bigdatau.org/search?query=machine+learning), allowing learners to explore and interact with the resources in ERuDIte. Second, we released the complete index in JSON-LD and N-Triples format at https://doi.org/10.5281/zenodo.1214375.

We applied our previous work on data exchange, Karma (Knoblock et al. 2012), to the learning resource, person, and organization records, which we internally store in a relational database, to generate RDF and JSON-LD data. Karma uses R2RML models to map tabulated data into RDF using semi-automated methods to speed up the mapping process. We applied Karma's framing mechanism to create collections of linked *schema:CreativeWork*, *schema:Organization*, and *schema:Person* entities as JSONL files.

5 Related Work

There has been a sustained interest in applying semantic web technologies to model educational resources, as seen by early vision papers, such as Bourda and Doan (2003) which proposed a semantic web for learning resources using an RDF schema version of the IEEE Learning Object Metadata standard, journal special issues (Anderson and Whitelock 2004), survey papers (Aroyo and Dicheva 2004; Dietze et al. 2013; Pereira et al. 2018), and the recent ISWC 2015 LINKed EDucation workshop. Other educational resource collection efforts include ELIXIR Training e-Support System (tess.elixir-europe.org), GOBLET (Brazas, Blackford, and Attwood 2017), and the Open Educational Resources Commons (oercommons.org). Our project shares many of the same goals of these previous efforts. Our focus has been on indexing recent video content in the rapidly evolving field of data science using automated methods for scalability. Moreover, in addition, to formalize the metadata of learning resources using semantic web standards and ontologies, we link persons and organizations to well-known sources, such as DBpedia and DBLP.

6 Discussion

We have presented ERuDIte, the BD2K TCC Educational Resource Discovery Index for Data Science. ERuDIte contains over 11,000 training resources on data science including courses (MOOCs), video tutorials, conference talks, technical books, and other materials. The metadata of these resources is described uniformly using Schema.org and automatically tagged with concepts from the Data Science Education Ontology (DSEO). To ensure high quality, the resources and metadata descriptions are curated by experts.

We represent ERuDIte as linked data to facilitate open access. We use record linkage techniques to map the references to people and organizations in the learning resource metadata to external entities in DBpedia, DBLP, and ORCID, thus embedding our collection in the web of linked data.

Our collection is continually growing. We are currently curating several thousand data science videos from YouTube and books from Google Books and plan to add them to our collection as they are reviewed and accepted. We hope that

Type	Source	Accuracy ($\geq \frac{2}{3}$)	Support ($\geq \frac{2}{3}$)	Accuracy (1)	Support (1)
Organization	DBpedia	0.816	98	0.905	84
Person	DBpedia	0.964	197	0.964	196
Person	DBLP	0.942	191	0.947	188

Table 5: Accuracy of automatic linking for test sets where $\geq \frac{2}{3}$ or all (1) of the annotators agree, and the number of matches in the test set (support).

ERuDIte will contribute to the growth of open linked educational resources on the web.

A major ongoing effort is to identify prerequisite relations between learning resources, e.g., that linear regression should be learned before logistic regression. We plan to provide personalized training paths using resource descriptions and prerequisite relations, as well as collected user interactions (searches, creation of educational plans, ratings) from the BigDataU.org web portal.

While ERuDIte is specifically focused on data science learning resources, our approaches to describe, enrich, and organize online learning resources could naturally extend to any domain with a collection of resources that needs structure to enable search and exploration. We hope that the approach of ERuDIte will encourage the creation of other enhanced resource aggregators, thus providing open collections and tools for continuous learning for anyone interested in delving into a specific knowledge domain.

Acknowledgments

This work is supported by the National Institutes of Health (NIH) under Grant 1U24ES026465. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of NIH. We thank the rest of the BD2K TCC team: John D. Van Horn, Jeana Kamdar, Sumiko Abe, Xiaoyu Lei, Avnish Bhattrai, and Crystal Stewart for their contributions to BigDataU.org.

References

Ambite, J. L.; Fierro, L.; Geigl, F.; Gordon, J.; Burns, G. A. P. C.; Lerman, K.; and Van Horn, J. D. 2017. BD2K ERuDIte: The educational resource discovery index for data science. In *Proc. of the International Conference on World Wide Web Companion*, 1203–11.

Anderson, T., and Whitelock, D. 2004. The educational semantic web: Visioning and practicing the future of education. *Journal of Interactive Media in Education* 1.

Aroyo, L., and Dicheva, D. 2004. The new challenges for e-learning: The educational semantic web. *Journal of Educational Technology & Society* 7(4):59–69.

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. DBpedia: A nucleus for a web of open data. In *The Semantic Web*, 722–735. Springer-Verlag.

Bourda, Y., and Doan, B. 2003. The semantic web for learning resources. In *Proc. of the IEEE International Conference on Advanced Learning Technologies*, ICALT, 322–3.

Brazas, M. D.; Blackford, S.; and Attwood, T. K. 2017. Training: Plug gap in essential bioinformatics skills. *Nature* 544(161).

Chen, D., and Manning, C. D. 2014. A fast and accurate dependency parser using neural networks. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 740–50.

Daiber, J.; Jakob, M.; Hokamp, C.; and Mendes, P. N. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proc. of the 9th International Conference on Semantic Systems* (*I-Semantics*). New York: ACM.

Dietze, S.; Sanchez-Alonso, S.; Ebner, H.; Yu, H. Q.; Giordano, D.; Marenzi, I.; and Bernardo, P. N. 2013. Interlinking educational resources and the web of data. *Program* 47(1):60–91.

Heath, T., and Bizer, C. 2011. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers.

Knoblock, C. A.; Szekely, P.; Ambite, J. L.; Goel, A.; Gupta, S.; Lerman, K.; Muslea, M.; Taheriyan, M.; and Mallick, P. 2012. Semi-automatically mapping structured sources into the semantic web. In *Proc. of the Extended Semantic Web Conference*.

Ley, M. 2002. The DBLP computer science bibliography: Evolution, research issues, perspectives. In *String Processing and Information Retrieval SPIRE*, 1–10.

McCallum, A. K. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu

Monge, A., and Elkan, C. 1997. An efficient domain-independent algorithm for detecting approximately duplicate database records. In *Proc. of the ACM SIGMOD workshop on Research issues in Data Mining and Knowledge Discovery*.

Naumann, F., and Herschel, M. 2010. *An Introduction to Duplicate Detection*. Morgan and Claypool Publishers.

Ohno-Machado, L. 2014. NIH's Big Data to Knowledge initiative and the advancement of biomedical informatics. *Journal of the American Medical Informatics Association* 193.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Pereira, C. K.; Siqueira, S.; Nunes, B. P.; and Dietze, S. 2018. Linked data in education: a survey and a synthesis of actual research and future challenges. *IEEE Transactions on Learning Technologies* 11(3):400–412.

Shahnaz, F.; Berry, M. W.; Pauca, V. P.; and Plemmons, R. J. 2006. Document clustering using nonnegative matrix factorization. *Information Processing & Management* 42(2):373–86.

Winkler, W. E. 1990. String comparator metrics and enhanced decision rule in the Fellegi–Sunter model of record linkage. In *Proc. of the Section on Survey Research Methods*, 354–9. American Statistical Association.

Winkler, W. E. 1999. The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Bureau of the Census.