

Content-based Comparison for Collections Identification

Weijia Xu¹ Ruizhu Huang¹ Maria Esteva¹ Jawon Song¹ Ramona Walls²

¹Texas Advanced Computing Center, University of Texas at Austin

²Cyverse.org

¹{xwj, rhuang, maria, jawon}@tacc.utexas.edu ²rwalls@cverse.org

Abstract— Assigning global unique persistent identifiers (GUPIs) to datasets has the goal of improving their accessibility and simplifying how they are referenced and reused. However, as repositories receive more and complex data, attesting for the identity of datasets attached to persistent identifiers over time is becoming more challenging. This is due to the nature of scientific research data, which is generated through distributed research practices and evolves across different computational environments. This work presents a robust, automated computational service for data content comparison as a valuable addition to assigning, managing, and tracking persistent identifiers. We operationalized the functions of the service within the archival space by linking data provenance and identity to authenticity. The need for such service is shown through three genomics data use cases in which the results aided curators establishing the identity of datasets and inferring issues of provenance. We describe the system's design, implementation and performance, and report on lessons learned.

Keywords— component; Data Curation; Content Based Comparison; Biological Sequences; Global Unique Persistent Identifiers; Spark

I. INTRODUCTION

Data driven research methods in the Sciences and the Humanities result in many new discoveries and in growing amounts of corresponding data. To preserve and make the data permanently accessible, most online repositories are providing curation services, which include granting global unique permanent identifiers (GUPIs) to the published datasets. However, while GUPIs are perceived as a synonym of permanence, many challenges remain in their implementation and maintenance [1]. Among those are: making decisions about assigning identifiers, attesting for the data identity over time, and continuously managing the provenance metadata needed for the data to be reused.

In today's digital environment such challenges are exacerbated by the distributed nature of scientific data practices and the growing sizes of the data. For example, repositories that assign GUPIs demand significant data preparation and manual description prior to publication, which can be daunting to accomplish for data curators. More often than not, such descriptions produce enough information for purposes of data citation but are insufficient for data reuse. As datasets are created and modified during the research process, they may result in multiple components and versions. Some data versions and components are submitted to canonical repositories, while others may be stored elsewhere and continue to be actively used by the

research team, deriving over time in new related data publications. Therefore, data from one project may exist in different formats, versions, and stages of completion, and while some transformations are documented, others may seem too trivial and thus easily ignored. All of this contributes to poor provenance information and failure to record post publication events and consequently the need to reaffirm the identity of a dataset in context with the related one.

Traditionally, problems such as the ones described above have been the concern of archival science and can be better understood in that space. In this project we operationalize the problems through the archival concepts of provenance and identity, and relate those to data authenticity. Provenance is understood as the processes through which data is created and modified over time, and identity as the possibility to distinguish data from other data through its content, its relations to other data, and contextual information recorded as metadata. However, when discussing digital data, there are differences with the notion of authentic archival records, which are preserved in an archive with all the information that attests to their authenticity. Those differences are related to how data is conceptualized by their creators and users. In the midst of often messy and distributed data practices, as new versions, copies, and derivatives of a dataset emerge, and where data is auto-published by its creators, reused beyond publication, and may be published again in different repositories and for different reasons, authenticity has to be continuously managed to clarify the roles and relations of the data.

In this environment, a service that can identify changes, connections, and differences between datasets is useful to maintain a record of the dataset's authenticity. And yet, although archivists have long defined the characteristics of authentic records, and those principles are gaining their way into the realm of data curation, they have not been developed into automated and scalable methods.

In particular, the biology community has expressed the need to better understand the use of identifiers across the research lifecycle for purposes of tracking provenance, and for assembling large and dispersed datasets [2]. As part of the Identifier Services (IDS) research project [3], we are investigating a content-based large data comparison framework. The IDS project is designed to track the evolution, integrity, and identity of biology datasets in relation to assigning, maintaining, and updating GUPIs. Through the content based comparison function, which is one of the project's micro-services, the differences between related data are computed. This function is relevant for very

large datasets such as sequencing files, whose contents cannot be manually compared.

We built a framework for genomics data comparison and applied it to three exemplary use cases. The service functions within IDS, and is external to the repositories where data are located. It uses AGAVE API [4] to transfer data to the high performance computing (HPC) resource Wrangler [5] where large-scale comparisons are performed. Through the results, curators can infer the role of the compared files and settle provenance and identity issues. Comparisons can be repeated over time to continuously manage the authenticity of evolving data.

II. BACKGROUND AND RELATED WORK

The archival construct of authenticity includes context, provenance, content, and integrity among the elements considered to attest that a record is what it claims to be over time and space. Clifford Lynch explains that determining authenticity in the digital environment needs to be managed over time, and that there are both mechanical and conceptual implications [6]. Mechanically, establishing authenticity for a digital object may involve verifying the integrity of the bits, its metadata, and its identifiers. Conceptually, an intelligent system could help determine and document series of related assertions about a digital object in connection to its provenance, versions, and derivatives [6].

Data authenticity cross references with data identity. Wynholds states that data identity is constructed through a combination of elements that aim to establish that an object is “definable and recognizable” from other objects [1]. The author points that in the current scientific environment data is an “unruly and poorly bounded object,” making it difficult to identify its uniqueness. This relates to the presence of multiple versions and or components of a dataset, which may be stored in different places at different times, and in many cases bearing different identifiers. Once an identical, similar, or related object appears, the identity of the data of interest can be challenged, all of which has an impact on the policies and tools surrounding the assignment of GUPIs.

However, repositories that are long-term custodians of data and grant unique identifiers have poor or no mechanisms to continuously assess identity beyond validating the integrity of the data they hold [3]. Most of today’s repositories are silos, and once an object is deposited, tracking differences and associations between data versions and or related components that are outside of the repository is either not practiced or is done manually by updating the metadata record. This landscape suggests the need for computational solutions to track, validate, and document data identity in a continuum and at scale.

A common approach used to verify and compare files in data repositories is through a hash function [7], [8]. A hash value is computed at the file bit level and, if two files being compared are identical, their hash value will match. While IDS allows conducting hash analysis for data integrity purposes, this is not sufficient to aid establishing identity. The method is not content-aware therefore, if the formats of the files to be compared are different, the hash value will be

different even if the content is identical. Furthermore, this approach cannot provide additional information regarding the nature of the differences between files, leaving users with limited options to make inferences about provenance.

For the biological sequence formats, which are at center of our investigation, a number of tools are available to provide more detailed comparisons [9], [10]. Those, however, are not ideal for establishing identity at the scale required. For example, the popular sequence comparison tool, BLAST [9] indexes one of the collections and then searches each record in the other collection against the index. This process is computationally expensive and generates extra information that is not needed for our purposes.

We introduce a powerful content based comparison framework. The comparison functions within the IDS application (<https://identifierservices.org>), where users can conduct different and complimentary data integrity and identity analyses to decide on the nature of the similarities and differences between related data.

III. SCALABLE COMPARISON FRAMEWORK

A. Workflow Overview

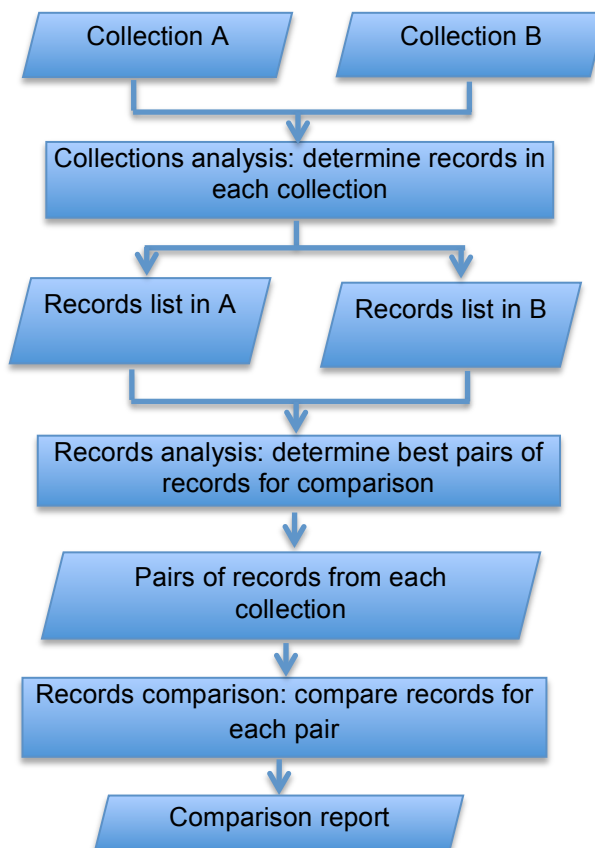


Figure 1. Workflow for comparing two collections.

Figure 1 shows an overview of the workflow for comparing two datasets. Here, a *collection* refers to a set of records as the analytic unit that will be compared. The first step of the workflow is *collections analysis* to determine the types of

records contained in the collections to be compared. Here, record types are inferred through available parsers that identify file formats and structures found in the genomics space. Note that a collection may have more than one type of record, as it may be composed by more than one file format. Once the parsers automatically identify the type of records to be compared the outputs are two lists, one for each collection, in which each record is represented as a <key, value> pair. The key is a unique identifier for the values of that pair within the collection. Next the *records analysis* step creates a list of record pairs from each collection for comparison. The identifiers are extracted for each file format and used as the key for each record; the value of each record is stored as a string object. During *records comparison*, the method of comparison between two records is selected based on the record types, and a summary report is created to show the differences between the collections.

B. Comparison algorithm and implementation

The basic steps of the comparison algorithm are outlined below.

1. Convert list of records as (identifier/key, value) pairs.
2. The records are sorted based on the identifiers.
3. Start with a, b as the pointer to the head of list A, B respectively according to a score function,
 - If $\text{score}(a, b) \leq 0$, record results and move a forward,
 - If $\text{score}(a, b) > 0$, record results and move b forward.

The algorithm sorts the lists of records based on their unique identifiers first and the comparison is carried out as linear traversal on both lists. The computational complexity is $O(n \cdot \log n)$, where n is the maximum number of records in each collection. This algorithm can also be parallelized on HPC resources for scalability and efficiency. Different distance functions can be used in step 3 to compare two records, and appropriate distance functions can be used for different record types. The comparison algorithm was designed to maximize comparison efficiency. We implemented standard string distance functions including: Hamming distance, prefix distance, suffix distance, and edit distance [11]. The collection analysis step used the BioPython package to recognize a variety of biological file formats such as: FASTA for gene sequences, FASTQ for short reads, and GFF for gene variation [12]. BioPython is a popular tool used in Bioinformatics and additional parsers can be added if needed. To handle large amounts of records, both the records analysis and the records comparison steps were implemented using the Spark processing framework [13], which can utilize parallel computing resources when available.

C. Results reporting

The results are categorized in four groups: 1) both keys and values match for both collections; 2) only keys match in both collections; 3) and 4) keys cannot be matched from one collection to the other (A to B, and B to A).

An important goal of this research is to convey information for users to understand the nature of the discrepancies between collections. From a clear understanding curators can decide how to assign or relate identifiers so that data can be recognized unequivocally in the midst of copies and versions. Moreover, they can infer the reasons of such differences and add provenance information to the research record. However, there are challenges reporting the results of large scale comparisons. For example, while all the differences between two collections can be recorded for users to review, this is neither useful nor intelligible, as the amount of differences identified by the comparison algorithm can be of the same order of the records being compared. Abstracting and representing comparison results in ways that allow understanding and decision-making is crucial to the usability of this computational service. After consulting with biologists, and curators about reporting requirements, the comparison report consists of three layers of information. The first layer is an overall statistical summary that includes the following information:

- Number of records identified and compared for each collection.
- Number of records successfully matched through their keys (id) between collections.
- Numbers of records that only exist in one of the collections (A or B).
- Histogram of the distribution scores between matched records.

The score between two records is normalized as a range between 0 and 1, with a score of 1 indicating identical records. The histogram is created using the scores of all non-identical pairs of records aggregated by every 10 percentiles.

The second layer of information is intended to help users understand the differences between collections. It includes examples of pairs of records randomly sampled from each collection, as well as records that exist only in one collection. The examples show the full record including the key value (id) and the record content.

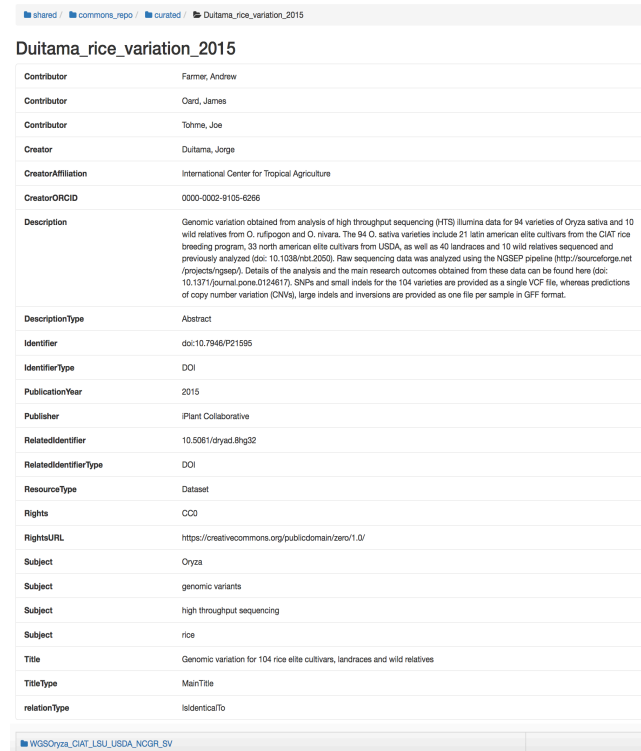
Although it may be too long for a user to inspect, the third layer is a complete copy of the comparison results. The results are organized based on matching status, i.e. identical records, non-identical records, and records missing in one of the collections. To reduce storage needs, only the unique identifier of each record is stored. This document functions as a registry and can be used to validate the comparison if the process has to be reproduced.

IV. USE CASES AND LESSONS LEARNED

Using the computational framework, we completed three use cases that represent different typical scenarios found in genomics data collections. We present each use case, the comparison results, and the decisions made by curators and researchers about the data identity.

A. Use case 1: Copies of a dataset are stored in different repositories with metadata and content discrepancies.

This case corresponds to a rice genomic variations dataset obtained from the analysis of high throughput sequencing (HTS) data for: 94 varieties of *Oryza Sativa*, and 10 wild relatives from *Oryza Rufipogon* and *Oryza Nivara* [14]. Details of the research and data outcomes can be found in [15]. SNPs (single nucleotide polymorphisms) and small indels (insertions and deletions) for the 104 varieties are provided as a single variance call file (VCF), whereas predictions of copy number variation (CNVs), large indels, and inversions are provided as one file per sample in genetic frequency file format (GFF). Biologists studying the genetic variances in plants frequently reference this dataset. Due to its popularity, multiple copies of the dataset exist in different repositories, two of which are the CyVerse Data Commons [16] and Dryad [17]. Both the metadata fields and the content values stored in the two repositories have noticeable differences. To view the Dryad metadata record go to: doi:10.5061/dryad.8hg32/1. To view the Cyverse metadata record go to: doi:10.7946/P21595.¹



Duitama_rice_variation_2015	
Contributor	Farmer, Andrew
Contributor	Oard, James
Contributor	Tohme, Joe
Creator	Duitama, Jorge
CreatorAffiliation	International Center for Tropical Agriculture
CreatorORCID	0000-0002-9105-6266
Description	Genomic variation obtained from analysis of high throughput sequencing (HTS) Illumina data for 94 varieties of <i>Oryza sativa</i> and 10 wild relatives from <i>O. rufipogon</i> and <i>O. nivara</i> . The 94 <i>O. sativa</i> varieties include 21 latin american elite cultivars from the CIAT rice breeding program, 33 north american elite cultivars from USDA, as well as 40 landraces and 10 wild relatives sequenced and previously analyzed (doi: 10.1038/nbt.2050). Raw sequencing data was analyzed using the NGSFP pipeline (http://sourceforge.net/projects/ngsfp/). Details of the analysis and the main research outcomes obtained from these data can be found here (doi: 10.1371/journal.pone.0124617). SNPs and small indels for the 104 varieties are provided as a single VCF file, whereas predictions of copy number variation (CNVs), large indels and inversions are provided as one file per sample in GFF format.
DescriptionType	Abstract
Identifier	doi:10.7946/P21595
IdentifierType	DOI
PublicationYear	2015
Publisher	iPlant Collaborative
RelatedIdentifier	10.5061/dryad.8hg32
RelatedIdentifierType	DOI
ResourceType	Dataset
Rights	CC0
RightsURL	https://creativecommons.org/publicdomain/zero/1.0/
Subject	Oryza
Subject	genomic variations
Subject	high throughput sequencing
Subject	rice
Title	Genomic variation for 104 rice elite cultivars, landraces and wild relatives
TitleType	MainTitle
relationType	IsIdenticalTo

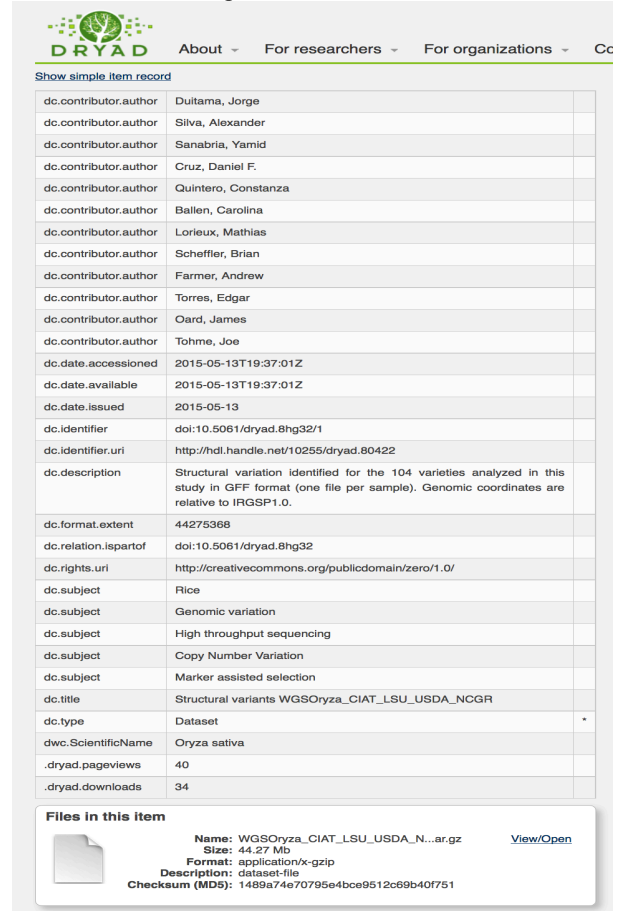
Figure 2 Metadata record for the Rice Genome Variation dataset in the Cyverse Data Commons repository

Each repository uses different fields to describe the same dataset, and the level of details entered in the corresponding fields are different as shown in Figure 2 and 3. Furthermore, the way in which data is delivered in both locations, and the functionalities of each repository platform

¹ The metadata in the Cyverse record has been edited to reflect the relationship with the dataset published in Dryad. However, the metadata record in Dryad was not edited to reflect the relationship.

differ. Dryad, an established data repository, which integrates data to journal publications, delivers the dataset as one compressed file. Instead, CyVerse provides access to all the files, and offers a direct connection to HPC resources to facilitate data analysis. The differences between both platforms explain why the researcher team wants to offer both access points to their data. Both Dryad and CyVerse assigned GUIs in the form of DOIs to each dataset. As a result, it was not straightforward for users to identify both datasets as being the same work. To clarify differences and similarities between the datasets and to complete the metadata record, the data curator at CyVerse needed more information.

We proceeded to compare the data in our framework by integrating the parser from the BioPython package to extract the list of records from each collection. The framework promptly verified that the two collections are exactly the same. Learning about the results, the data curator at CyVerse clarified the connection between the datasets by relating the Dryad identifier as an identical dataset in the metadata of the CyVerse identifier (See full metadata record at <http://ezid.cdlib.org/id/doi:10.7946/P21595>). In this case, both datasets are authentic, but it was important to clarify the roles and relationships between them.




DRYAD About - For researchers - For organizations - Cc	
Show simple item record	
dc.contributor.author	Duitama, Jorge
dc.contributor.author	Silva, Alexander
dc.contributor.author	Sanabria, Yamid
dc.contributor.author	Cruz, Daniel F.
dc.contributor.author	Quintero, Constanza
dc.contributor.author	Ballen, Carolina
dc.contributor.author	Lorieux, Mathias
dc.contributor.author	Scheffler, Brian
dc.contributor.author	Farmer, Andrew
dc.contributor.author	Torres, Edgar
dc.contributor.author	Oard, James
dc.contributor.author	Tohme, Joe
dc.date.accessioned	2015-05-13T19:37:01Z
dc.date.available	2015-05-13T19:37:01Z
dc.date.issued	2015-05-13
dc.identifier	doi:10.5061/dryad.8hg32/1
dc.identifier.uri	http://hdl.handle.net/10255/dryad.80422
dc.description	Structural variation identified for the 104 varieties analyzed in this study in GFF format (one file per sample). Genomic coordinates are relative to IRGSP1.0.
dc.format.extent	44275368
dc.relation.ispartof	doi:10.5061/dryad.8hg32
dc.rights.uri	http://creativecommons.org/publicdomain/zero/1.0/
dc.subject	Rice
dc.subject	Genomic variation
dc.subject	High throughput sequencing
dc.subject	Copy Number Variation
dc.subject	Marker assisted selection
dc.title	Structural variants WGSOrzya_CIA_T_LSU_USDA_NCGR
dc.type	Dataset
dwc.ScientificName	<i>Oryza sativa</i>
.dryad.pageviews	40
.dryad.downloads	34
Files in this item	
	Name: WGSOrzya_CIA_T_LSU_USDA_N...ar.gz View/Open
	Size: 44.27 MB
	Format: application/x-gzip
	Description: dataset-file
	Checksum (MD5): 1489a74e70795e4bce9512c69b40f751

Figure 3 Metadata record for the for Rice Genome Variation in Dryad

B. Use case 2: Two components of a dataset are published in different repositories and show differences.

The second use case is a common scenario found in the process of publishing a dataset. A research group wants to publish a complete dataset resulting from the analysis of five different maize lines. The dataset includes the FASTQ files, the BAM files, and the 100BP files corresponding to the outcomes of the sequencing, alignment, and analyses of methylation processes of each maize line. Component parts of this dataset, the raw sequencing files in FASTQ format, are already published in the canonical Sequence Read Archive (SRA) repository which accepts only sequencing files [18]. In turn, the working copy of the entire dataset is stored at the Texas Advanced Computing Center (TACC) for active analysis, and the team wants to make it public for others to reuse all the component parts. They also want to link the complete set to journal publications in which they use more than the sequencing files.

Before assigning a GUI and publishing the complete dataset, the curators wanted to know if the raw sequencing files stored at TACC were identical to those already published in SRA. After conducting the content based comparison, the results showed that despite the researchers' belief, the FASTQ files stored in the SRA repository were different from those stored at TACC. The researchers needed to see the results and evaluate the nature of the discrepancies.

Table 1. Example of comparison summary between the researchers working copy and the archived data with SRA

1. Records matched in both collections	165,993,413
2. Identical records in both collections	23,222,181
3. Different records in both collections	142,771,232
4. Records found only in A	0
5. Records found only in B	1,769,469

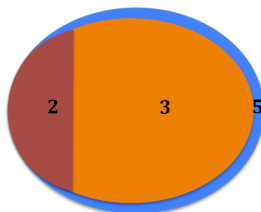


Figure 4. Set diagram of comparisons summary in Table 1

Table 1 shows that the raw sequencing files in the complete set at TACC are different from those in the SRA repository. Given the extensive amount of records that were compared, we offered the results as percentages of differences between both collections. Along with the results, we presented records as examples of those differences. The working copy of raw sequencing file (set A) has 165,993,413 sequences, and the copy from SRA (set B) has 167,762,882 sequences. All IDs in A can be found in B and within these common IDs there are 23,222,181 common sequences shown as '2' in Figure 4. Observing the results, the research team concluded that the percentage of difference between

both collections was not significant. Importantly, by looking at the sample records offered as the second layer of information, the team understood the nature of the discrepancies. They realized that the data stored at TACC had been processed by a common technique known as adaptive trimming, and that the untrimmed raw files were stored at SRA. Still, researchers considered that both trimmed and un-trimmed collections could be considered the same "work". They concluded that a DOI could be assigned to the complete dataset at TACC. The curator will relate the SRA identifier in the DOI DataCite metadata to clarify that both are authentic, what differences they have, and the reason why there are two datasets (complete and component) published.

C. Use case 3: Two datasets with similar metadata show significant content differences.

The third use case involves two datasets that share almost the same metadata record but the content of the data is significantly different. We identified two datasets, DDEV [19] and IFCJ [20] in the 1K plant genome project [21]. For both datasets, 12 out of 14 metadata fields available through the web API are identical. The differences are in the "RNA extractor" and "note" fields (Figure 5).

Details for sample code #DDEV		Details for sample code #IFCJ	
ID:	DDEV	ID:	IFCJ
Clade:	Magnoliids	Clade:	Magnoliids
Order:	Canellales	Order:	Canellales
Family:	Canellaceae	Family:	Canellaceae
Species:	Canella winterana	Species:	Canella winterana
Tissue Type:	young leaves	Tissue Type:	young leaves
Status:	sequenced	Status:	sequenced
Voucher Data:	Soltis and Miles 2995	Voucher Data:	Soltis and Miles 2995
Sample Preparation:	GenVault	Sample Preparation:	GenVault
Sample Provider:	D. Soltis	Sample Provider:	D. Soltis
Sample Shipping Contact:	I. Jordon-Thaden	Sample Shipping Contact:	I. Jordon-Thaden
RNA Extractor:	I. Jordon-Thaden	RNA Extractor:	D. Soltis
RNA Shipping Contact:	I. Jordon-Thaden	RNA Shipping Contact:	I. Jordon-Thaden
Subsets:	Angiosperm	Subsets:	Angiosperm
Notes:	also shipped Soltis-2011Nov / GH	Notes:	

Figure 5. Screenshots of the metadata records of the two datasets at the 1K plant genome website.

Despite the high similarity among the metadata records, the comparison results indicated that the datasets are different. The number of records between them differ; one dataset has 13,350,236 records and the other one 18,500,835. Figure 6 below shows the score distribution of the comparison results.

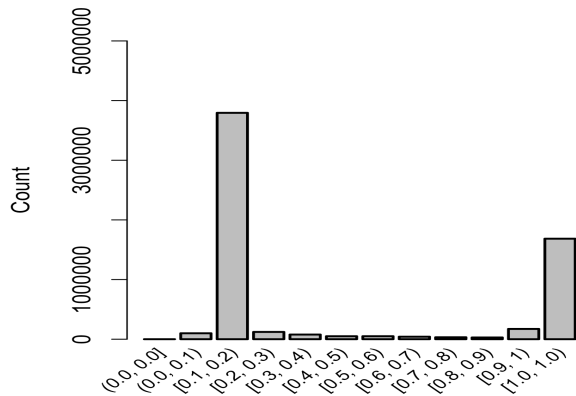


Figure 6. Distribution of matching scores between two datasets from the 1K Plant Genome Project.

Figure 6 shows that most records in both datasets are different. However, we also notice a peak of scores between 0.1 and 0.2. In this case we used the prefix-matching algorithm as a distance function. The score distribution indicates that a relatively large quantity of record pairs match in the first 20% segment of the records being compared

Previously to the comparison, the researchers of the 1K Plant Genome Project mentioned that the differences between the two datasets could be due to the contamination of one of the plant samples, but they still wanted to publish both datasets. After observing the results, they suggested that the pair of matches during the first 20% segment of the records could be due to the use of the same sequencing primer in both datasets. Additionally, they did not expect to find such similarity between both collections, and are interested in further investigating the results. The comparison results improved their knowledge about the data provenance and allowed making further curatorial decisions. Concluding that the existing metadata is insufficient for users to determine the differences and thus the identity of each dataset, the curators indicated that each published dataset needs its own unique identifier as well as clarifications about their provenance in the metadata record. Again, the researchers decided that they wanted to maintain both datasets publicly available for users to explore.

D. Performance and scalability

As datasets become increasingly large, the content-based comparisons can be very computationally expensive. Hence, an important requirement of content-based data comparison is that it is feasible and scalable. We conducted tests to determine the scalability and performance of the framework.

All the computations were conducted in the data intensive system Wrangler [5]. Figure 7 shows the execution time involved in comparing two sequencing files in use case 3. The three main steps during the comparison shown in Figure 7 are: 1) time spent reading the data, 2) time spent matching the records, and 3) computing the scores between pairs of records. The two files compared are ~3GB each, with 13,350,236 and 18,500,835 records respectively.

To evaluate performance we conducted the comparisons using different number of nodes (Figure 7). The blue columns show job runs using four nodes and the red ones show job runs using eight nodes.

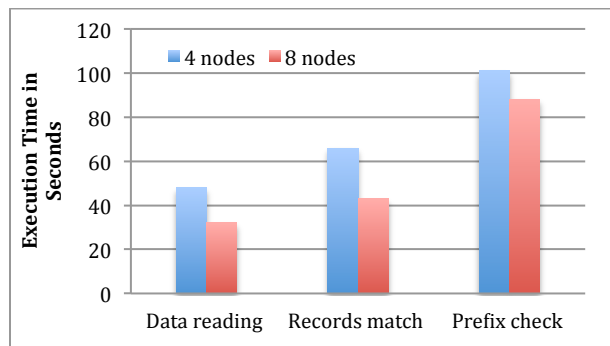


Figure 7. Performance and scalability of the main content-based comparison steps.

Our implementation performed the comparison in under two minutes. We observed that the performance increased only 1/3rd so it is not linearly correlated to the number of nodes used. The reason for this is that the comparison is largely dominated by the read data operation, which is less scalable, regardless of the number of nodes used. If the data increases, the calculation can be parallelized even further to improve scalability. In addition we estimated that running the comparisons using the existing sequence comparison tool, Blast would takes hours instead of minutes.

V. CONCLUSIONS

In distributed research and computational environments, as copies and versions of a dataset are generated, the identity and provenance of data has to be managed over time to continuously support a dataset's authenticity in relation to other objects that exist and evolve. In the context of the IDS research project, we created and tested a content-based comparison framework to aid curators determine the uniqueness of a dataset. The three use cases described in this paper are common scenarios found in genomics data management. In all cases, the metadata alone did not allow establishing their unique identity, the researchers were not quite sure of what had changed in the data, and assigning identifiers was not clear. Conducting content-based comparison between the datasets was useful to make identity decisions, provision GUPs, and facilitate documenting provenance. The way in which the comparison results are presented help curators understand the reasons for the differences or similarities between the compared data. The work ahead involves generalizing the framework to other data types .

ACKNOWLEDGMENT

This work is supported by the NSF grants #26100741 (Evaluating Identifier Services for the Lifecycle of Biological Data) and #1341711 (Wrangler: A

Transformational Data Intensive Resource for the Open Science Community).

REFERENCES

- [1] L. Wynholds, "Linking to scientific data: Identity problems of unruly and poorly bounded digital objects," *International Journal of Digital Curation*, vol. 6, no. 1, pp. 214–225, 2011.
- [2] R. P. Guralnick *et al.*, "Community next steps for making globally unique identifiers work for biocollections data," *ZooKeys*, no. 494, pp. 133–154, Apr. 2015.
- [3] M. Factor, E. Henis, D. Naor, S. Rabinovici-cohen, P. Reshef, and S. Ronen, "Authenticity and Provenance in Long Term Digital Preservation: Modeling and Implementation in Preservation Aware Storage," *Society*, pp. 1–10, 2009.
- [4] "Agava Platform." [Online]. Available: <https://agaveapi.co/>. [Accessed: 04-Nov-2016].
- [5] C. Jordan, D. Walling, W. Xu, S. A. Mock, N. Gaffney, and D. Stanzione, "Wrangler's user environment: A software framework for management of data-intensive computing system," in *Big Data (Big Data), 2015 IEEE International Conference on*, 2015, pp. 2479–2486.
- [6] C. Lynch, "Authenticity and integrity in the digital environment: an exploratory analysis of the central role of trust," *Authenticity in a Digital Environment*, pp. 32–50, 2000.
- [7] E. Grosse, "Repository mirroring," *ACM Transactions on Mathematical Software*, vol. 21, no. 1, pp. 89–97, 1995.
- [8] G. Sivathanu, C. P. Wright, and E. Zadok, "Ensuring Data Integrity in Storage: Techniques and Applications," *Proceedings of the ACM workshop on Storage security and survivability*, pp. 26–36, 2005.
- [9] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–10, 1990.
- [10] H. Li and N. Homer, "A survey of sequence alignment algorithms for next-generation sequencing," *Briefings in Bioinformatics*, vol. 11, no. 5, pp. 473–483, 11-May-2010.
- [11] D. Gusfield, G. M. Landau, and B. Schieber, "An efficient algorithm for the All Pairs Suffix-Prefix Problem," *Information Processing Letters*, vol. 41, no. 4, pp. 181–185, Mar. 1992.
- [12] P. J. A. Cock *et al.*, "Biopython: Freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, Jun. 2009.
- [13] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster Computing with Working Sets," in *HotCloud'10 Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, 2010, p. 10.
- [14] X. Xu *et al.*, "Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes," *Nature biotechnology*, vol. 30, no. 1, pp. 105–11, 2012.
- [15] J. Duitama *et al.*, "Whole genome sequencing of elite rice cultivars as a comprehensive information resource for marker assisted selection," *PLoS ONE*, vol. 10, no. 4, Apr. 2015.
- [16] "Duitama_rice_variation_2015." [Online]. Available: <http://doi.org/10.7946/P21595>. [Accessed: 14-Nov-2016].
- [17] "Structural variants WGSORyza_CIAAT_LSU_USDA_NCGR." [Online]. Available: <http://datadryad.org/resource/doi:10.5061/dryad.8hg32/1>. [Accessed: 14-Nov-2016].
- [18] Q. Li *et al.*, "Examining the Causes and Consequences of Context-Specific Differential DNA Methylation in Maize," *Plant Physiology*, vol. 168, no. 4, pp. 1262–1274, 2015.
- [19] "Details for sample code #DDEV." [Online]. Available: <http://www.onekp.com/samples/single.php?id=DDEV>. [Accessed: 14-Nov-2016].
- [20] "Details for sample code #IFCJ." [Online]. Available: <http://www.onekp.com/samples/single.php?id=IFCJ>. [Accessed: 14-Nov-2016].
- [21] N. Matasci *et al.*, "Data access for the 1,000 Plants (1KP) project.," *GigaScience*, vol. 3, no. 1, p. 17, 2014.