# An Assessment of VTechData with respect to the CoreTrustSeal Repository Certification Requirements

## February 2019


## Virginia Tech University Libraries

Naina Pisharoti

Jonathan Petters

**Interviewees**

**Digital Imaging and Preservation Services**

Nathan Hall

Alex Kinnaman

Luke Menzies

**Information Technology Services**

Aaron Hunnewell

Lee Hunter

Jim Tuttle

**Data Services**

R. Shane Coleman

Andi Ogier

Jonathan Petters

## Summary

This report provides a brief internal assessment as to how well the current status of VTechData ("The Repository) meets the CoreTrustSeal repository certification requirements. A relatively lightweight process that is gaining popularity in research data management circles, this internal audit was carried out as an initial step towards the CoreTrustSeal Certification.

CoreTrustSeal has a set of 16 requirements that need to be fulfilled in order to qualify for the certification. Each requirement further has a level of compliance that needs to be indicated and is useful to give a quick overview of the status of the requirement. Data Services (DS), Digital Imaging and Preservation Services (DIPS) and Information Technology Services (ITS) were interviewed towards this report and assessment of VTechData with respect to 15 of the 16 requirements. The requirement regarding funding and staffing was deemed out of scope for this report.

VTechData has seen substantial development and improvements from when the first dataset was published in May of 2016 to the present day. However, this internal assessment shows that VTechData does not yet meet the CoreTrustSeal certification requirements.

CoreTrustSeal certification can be obtained if all requirements are either rated at a 4 (fully implemented) or a high 3 (well on the way towards implementation). As seen on the Table directly following this summary, the internal assessment of the Repository resulted in

- four ratings of 2 (The Repository has a theoretical concept)
- six ratings of 3 (The Repository is in the implementation phase)
- two ratings of 4 (The guideline has been fully implemented in the Repository)
- three ratings of 0 (Not applicable)
- one rating not evaluated (beyond the scope of this report)

While the Repository has seen at least some development towards meeting every certification requirement, the degree of implementation towards each requirement varies widely. Additionally, four of the requirements were deemed as not applicable to an institutional repository like VTechData, but improvements can still be made to help the Repository address these requirements.

Data Services and Information Technology Services (under which Digital Imaging and Preservation Services is now organized) will continue to work together to improve VTechData in regards to to the CoreTrustSeal requirements.

| No | Requirement Description | Rating | Recommendations (Responsible Entities) |
|---|---|---|---|
| 1 | **The repository has an explicit mission to provide access to and preserve data in its domain.** | 4 | • The Repository would need documentation of approval and support for the Repository within the organization for the CoreTrust Seal certification process. (DS) |
| 2 | **The repository maintains all applicable licenses covering data access and use and monitors compliance.** | 3 | • The Repository should run Deposit and Publishing agreements, Terms of Use, and non-compliance procedures by General Counsel for their approval. (DS)<br>• The Repository should document the procedure for its actions when deposited data violates the Deposit and/or Publishing Agreement. (DS) |
| 3 | **The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.** | 2 | • The Repository should continue to develop and document measures taken to ensure access to Repository holdings. (ITS)<br>• The Repository should continue to develop and document a succession plan for the Repository's data holdings to ensure their availability in the future, including roles, responsibilities, and a defined preservation period. (ITS, DS, DIPS) |
| 4 | **The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.** | 3 | • The Repository should standardize and document its communication procedure with depositors and publishers regarding confidentiality and ethics requirements. (DS)<br>• The Repository should create 'click through' Deposit and Publishing agreements, or at least put the policies up front for depositors and publishers prior to them interacting with the Repository platform. (DS, ITS) |
| 5 | **The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.** | - | Evaluation of The Repository toward this requirement was deemed out of scope for this report. |
| 6 | **The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, including scientific guidance, if relevant).** | 2 | • The Repository should convene an advisory committee of faculty members at Virginia Tech on how the Repository should develop to better meet the needs of the Virginia Tech research community. (DS) |
| 7 | **The repository guarantees the integrity and authenticity of the data.** | 2 | • The Repository should build the preservation infrastructure as planned by DIPS and integrate the Repository submission and dissemination platform within this preservation infrastructure. (DS, DIPS, ITS)<br>• The Repository should establish clear roles for each department regarding data integrity and authenticity, and improve coordination among all three entities. (DS, DIPS, ITS) |
| 8 | **The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.** | 0 | • The Repository could leverage external consulting services such as Springer/Nature's Research Data Support or participate in the Data Curation Network project when that becomes possible. In some cases Data Services may also be able to leverage their informatics consultants for some modicum of assistance with relevance and understandability assessment. (DS) |

| | | | |
|---|---|---|---|
| **9** | **The repository applies documented processes and procedures in managing archival storage of the data.** | 3 | • The Repository should replace the current hardware used, either with new hardware or through migration to cloud services. (ITS)<br>• The Repository should hold regular meetings to gain better insight into the operation of VTArchive and to better coordinate preservation actions. (DIPS, ITS) |
| **10** | **The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.** | 3 | • The Repository should continue working together to develop a workflow for effectively bagging and preserving Repository content. (DS, DIPS) |
| **11** | **The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.** | 0 | • The Repository could convene a team of disciplinary experts to help assess the quality of the content in the dataset, and leverage external efforts to develop guidance for assessing dataset fitness for use (e.g. the work of the WDS/RDA Assessment of Data Fitness for Use WG). (DS, See Requirement 8) |
| **12** | **Archiving takes place according to defined workflows from ingest to dissemination.** | 2 | • The Repository should improve detail of documentation for all workflows, and ideally such that new personnel could execute these workflows without further guidance. (DS, DIPS, ITS)<br>• The Repository should automate the portions of the ingest, publication and preservation workflow that can be automated. (DS, DIPS, ITS) |
| **13** | **The repository enables users to discover the data and refer to them in a persistent way through proper citation.** | 4 | • The Repository should allow metadata to be machine accessible directly through the Repository Platform. (DS, ITS) |
| **14** | **The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.** | 0 | • The Repository could continue developing a preservation metadata server and create/convert all preservation metadata into RDF. (DIPS, ITS) |
| **15** | **The repository functions on well-supported operating systems and other core infrastructural software and is using hardware** | 3 | • The Repository should complete documentation of its technical infrastructure. (ITS)<br>• The Repository should move away from usage of the Samvera platform for ingest and dissemination of content. (DS, ITS) |
| **16** | **The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.** | 3 | • The Repository should complete documentation of its technical infrastructure. (ITS)<br>• The Repository should should formalize its succession plan and data preservation plans and workflows. (DS, DIPS, ITS)<br>• The Repository should clarify responsibilities and roles regarding security. (DIPS, ITS) |

| **Rating** | **Description** |
|---|---|
| 0 | Not Applicable |
| 1 | The Repository has not considered this yet |
| 2 | The Repository has a theoretical concept |
| 3 | The Repository is in the Implementation Phase |
| 4 | The guideline has been fully implemented in the Repository |

# CONTENTS

# Introduction

VTechData is an Institutional Repository which focuses on collecting and preserving the data generated by researchers at Virginia Tech as part of its land-grant mission. An Institutional Repository can be defined as an archive which stores, preserves and provides access to physical or digital information created by users who are part of an institution, particularly a research institution.

VTechData (aka "The Repository" throughout this report) operations and development involves three primary entities in the Virginia Tech Libraries - Data Services **(DS)**, Digital Imaging and Preservation Services **(DIPS)** and Information Technology Services **(ITS)**[1]. DS handles the administrative and curatorial duties while DIPS and ITS provide preservation and system support. All three departments collaborate to ensure that ingest (deposit), dissemination and preservation services are provided to Virginia Tech researchers.

The purpose of this report is to provide a brief internal assessment as to how well the current status of VTechData meets the CoreTrustSeal repository certification requirements. A relatively lightweight process that is gaining popularity in research data management circles, this internal audit was carried out as an initial step towards the CoreTrustSeal Certification. If awarded, this certification would help us demonstrate credibility and trustworthiness in VTechData.

CoreTrustSeal has a set of 16 requirements that need to be fulfilled in order to qualify for the certification. Each requirement further has a level of compliance that needs to be indicated and is useful to give a quick overview of the status of the requirement.

Levels of Compliance (Rating) for each CoreTrustSeal Requirement:

0 – Not applicable
1 – The Repository has not considered this yet
2 – The Repository has a theoretical concept
3 – The Repository is in the implementation phase
4 – The guideline has been fully implemented in the Repository

In order to conduct this self audit, DS, DIPS and ITS personnel (i.e. the Interviewees on the title page) were interviewed based on the requirements relevant to their role. On the basis of what each department had to say, the authors awarded a rating of the current state of the repository, and made recommendations that would help VTechData ("The Repository") improve towards and/or meet its level of compliance with the CoreTrustSeal requirements.

[1] The bulk of this report was completed prior to DIPS being reorganized under ITS, and thus the two entities are kept separate for the purposes of this report.

The report is divided into fifteen parts, each part giving the description, assessment, recommendation and rating for each of the fifteen of the sixteen requirements (skipping Requirement 5 on funding and staffing). Within these requirements the term "Designated Community" is used at multiple instances: the Designated Community of VTechData comprises researchers at Virginia Tech who can deposit and publish their data therein.

VTechData has seen substantial development and improvements from when the first dataset was published in May of 2016 to the present day. However, this internal assessment shows that VTechData does not yet meet the CoreTrustSeal certification requirements.

While the Repository has seen at least some development towards meeting every certification requirement, the degree of implementation towards each requirement varies widely. Additionally, three of the requirements (8, 11 and 14, see table on pages 2 and 3) were deemed as not applicable to an institutional repository like VTechData, but improvements can still be made to help the Repository address these requirements.

Data Services and Information Technology Services (under which Digital Imaging and Preservation Services is now organized) will continue to work together to improve VTechData as compared to the CoreTrustSeal requirements.

## Requirement 1 (Mission Statement)

**Requirement Description:**

**The repository has an explicit mission to provide access to and preserve data in its domain.** *Repositories take responsibility for stewardship of digital objects, and to ensure that materials are held in the appropriate environment for appropriate periods of time. Depositors and users must be clear that preservation of, and continued access to, the data is an explicit role of the repository.*

**Assessment:**

The Repository VTechData is an institutional repository hosted by Virginia Tech that focuses on collecting and preserving research data generated at the University. The mission statement of the repository given on the online Repository platform clearly states the following:

"*The purpose of VTechData is to highlight, preserve, and provide access to research products (e.g. datasets) of the Virginia Tech community, and in doing so help to disseminate the intellectual output of the university in its land-grant mission. VTechData and Virginia Tech serve the Commonwealth of Virginia, the nation, and the world's community through the discovery and dissemination of new knowledge."*

VTechData is managed by three entities in the Virginia Tech University Libraries. It is a collaborative effort between Data Services, Digital Imaging and Preservation Services and Information Technology Services.

The repository and its mission is well established and has the full support of Libraries administration, i.e. the Dean of Libraries and the Associate Dean of Research and Learning.

**Recommendations for Repository to meet/improve compliance with this Requirement:**

- The Repository would need documentation of approval and support for the Repository within the organization for the CoreTrust Seal certification process (e.g. a short letter of support). (DS)

**Rating: 4 (The guideline has been fully implemented in the repository):**
*The mission statement is clearly defined and understood by all the entities involved in the functioning of VTechData. The Repository has full support of the higher-level administration.*

## Requirement 2 (Licenses)

**Requirement Description:**

**The repository maintains all applicable licenses covering data access and use and monitors compliance.**

*Repositories must maintain all applicable licenses covering data access and use, communicate about them with users, and monitor compliance. This Requirement relates to the access regulations and applicable licenses set by the data repository itself, as well as any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information.*

**Assessment:**

The data published on VTechData is openly accessible and The Repository policy prohibits any form of sensitive data. The deposit process within VTechData is tailored such that users get to choose a license for reuse and redistribution of their published data. Currently the Creative Commons 3.0 license (ex. Creative Commons 3.0 Attribution License) suite is allowed, with 'all rights reserved', 'CC0 1.0 Universal' and 'Public Domain Mark 1.0' as additional options.

As mentioned in the Terms of Use, VTechData users are required to abide by the policies enforced by Virginia Tech's Policy 7000: Acceptable Use and Administration of Computer and Communication Systems and other University Policies. The Policies page also gives a detailed description of the Deposit and Publication Agreements.

The Deposit Agreement gives the depositor a clear description of the levels of visibility to which they can restrict data access, but all published data is required to be openly accessible. It also strictly states that the repository does not host any form of data that

- Violates relevant University Policies,
- That could lead to direct or indirect identification of human subjects/students,
- That did not receive consent for sharing from human subjects or are in violation of IRB protocols, and
- Violate federal or state laws (HIPPA, FERPA, ITAR etc.).

The VTechData administrators have the responsibility of ensuring compliance with VT Policy 7030 (Policy on Privacy Statements on Virginia Tech Web Sites) to maintain the privacy of its users. Administrators also hold the rights to remove any data from the repository that do not comply with the requirements mentioned above.

Currently there is no documented procedure in place to identify and take action against any deposited data that violates the Deposit or Publishing agreements. Although, DS states that they would send an e-mail to the depositor stating that the data was non-compliant to the agreement and take the data down.

**Recommendations for Repository to meet/improve compliance with this Requirement:**

- The Repository should run Deposit and Publishing agreements, Terms of Use, and non-compliance procedures by General Counsel for their approval. (DS)
- The Repository should document the procedure for its actions when deposited data violates the Deposit and/or Publishing Agreement. (DS)

**Rating: 3 (in implementation phase):**
*Repository has established licensing for reuse and redistribution of datasets and conditions of platform use but has no procedure for identifying and acting upon deposited data that violates the deposit agreement. The latter can be implemented in a relatively short time period.*

## Requirement 3 (Continuity of Access)

**Requirement Description:**

**The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.**
*This Requirement covers the measures in place to ensure access to, and availability of, data holdings, both currently and in the future. Reviewers are seeking evidence that preparations are in place to address the risks inherent in changing circumstances.*

**Assessment:**

The ongoing access and preservation of The Repository is handled by the Digital Imaging and Preservation Services (DIPS) in collaboration with Information Technology Services (ITS).

DIPS aims to ensure all digital Libraries content (including Repository holdings) can be accessed and used in the future. Further objectives and policies have been detailed on their wiki. In order to maintain the current data holdings, regular checksums and fixity checks are to be run on all Libraries content. DIPS aims to preserve Libraries content for a period of at least 5-10 years.

The Repository currently does not have a continuity and succession plan. ITS is currently drafting a 3-year strategic plan which involves developing measures to handle catastrophic losses of data, which may include:

- a documented preservation plan covering digital objects, application code, server configurations (e.g., Ansible playbooks)
- a documented plan for restoration of digital objects from local backups and preservation services
- regularly scheduled testing of restoration

- regular testing of fail-over hardware (depending on the uptime expectations for the service)

Roles and responsibilities for Repository holdings is unclear among the departments involved and there is not yet agreement in terms of preservation periods.

**Recommendations for Repository to meet/improve compliance with this Requirement:**

- The Repository should continue to develop and document measures taken to ensure access to Repository holdings. (ITS)
- The Repository should continue to develop and document a succession plan for the Repository's data holdings to ensure their availability in the future, including roles, responsibilities, and a defined preservation period. (ITS, DS, DIPS)

**Rating: 2 (The repository has a theoretical concept):**
*The documentation/continuity plan about how the Libraries will provide for access and preservation of VTechData holdings is still under development. The role of each department needs to be clearly defined and agreed upon.*

## Requirement 4 (Confidentiality/Ethics)

**Requirement Description:**

**The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.**
*Adherence to ethical norms is critical to responsible science. Disclosure risk—for example, the risk that an individual who participated in a survey can be identified or that the precise location of an endangered species can be pinpointed—is a concern that many repositories must address. Evidence sought is concerned with not only having good practices for data with disclosure risks, but also the necessity to maintain the trust of those agreeing to have personal/sensitive data stored in the repository.*

**Assessment:**

DS, who administer VTechData state that this requirement is not fully applicable to the repository as they do not accept or host any form of sensitive data. The Deposit Agreement clearly explains that any datasets that do not de-identify human subjects' data in agreement with IRB protocols, are not in compliance with any of the federal, state and University policies, or have been published by violating copyright agreements will not be hosted.

Although these confidentiality and ethical requirements are mentioned in the Deposit Agreement, there is no standardized documented procedure to enforce them. DS do not currently have an efficient way of checking for identifiable and sensitive information in every format that research data is deposited into The Repository. They generally adopt an ad-hoc process while curating; that is, depending on the type of dataset, certain questions are asked to ensure the depositors have stuck to ethical norms and checks for sensitive content are made where feasible.

**Recommendations for Repository to meet/improve compliance with this Requirement:**

- The Repository should standardize and document its communication procedure with depositors and publishers regarding confidentiality and ethics requirements. (DS)
- The Repository should create 'click through' Deposit and Publishing agreements, or at least put the policies up front for depositors and publishers prior to them interacting with the Repository platform. (DS, ITS)

**Rating: 3 (The Repository is in the Implementation Phase):**

*While the Repository currently does not accept sensitive data and does not plan to in the near future, the Repository does have policies in place to inform depositors of the kinds of data that are acceptable and will work to better maintain ethical norms for Repository usage.*

## Requirement 5 (Funding and Staffing)

**Requirement Description:**

**R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.**

*Repositories need funding to carry out their responsibilities, along with a competent staff who have expertise in data archiving.*

**Evaluation of The Repository toward this requirement was deemed out of scope for this report.**

## Requirement 6 (Expert Guidance)

**Requirement Description:**

**The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, including scientific guidance, if relevant).**

*An effective repository strives to accommodate evolutions in data types, data volumes, and data rates, as well as to adopt the most effective new technologies in order to remain valuable to its Designated Community. Given the rapid pace of change in the research data environment, it is therefore advisable for a repository to secure the advice and feedback of expert users on a regular basis to ensure its continued relevance and improvement.*

**Assessment:**

The Repository does not have a formal procedure in place for seeking expert advice from its Designated Community (i.e. Virginia Tech researchers). Modifications to repository capabilities are considered upon feedback from Repository depositors and publishers on more of an ad-hoc basis. DS does record feedback from the Designated Community in user stories that help in evaluating and tracking common Repository issues.

Currently, there is no formal committee or panel to seek advice, thereby, DS relies mostly on their own expertise. The Samvera community is sometimes consulted for queries regarding the platform. Some other external communities such as DataCure and RDAP (Research Data Access and Preservation) are occasionally consulted as well.

**Recommendations for Repository to meet/improve compliance with this Requirement:**

- The Repository should convene an advisory committee of faculty members at Virginia Tech on how the Repository should develop to better meet the needs of the Virginia Tech research community. (DS) This advisory committee could be drawn from faculty members who are aware of and meet data publishing requirements of reputable research journals.

**Rating: 2 (The repository has a theoretical concept):**
*The Repository currently relies on their own expertise and access to other experts, but does not have regular input from its Designated Community (i.e. Virginia Tech Researchers).*

## Requirement 7 (Data Integrity and Authenticity)

**The repository guarantees the integrity and authenticity of the data.**
*The repository should provide evidence to show that it operates a data and metadata management system suitable for ensuring integrity and authenticity during the processes of ingest, archival storage, and data access.*

*Integrity ensures that changes to data and metadata are documented and can be traced to the rationale and originator of the change.*
*Authenticity covers the degree of reliability of the original deposited data and its provenance, including the relationship between the original data and that disseminated, and whether or not existing relationships between datasets and/or metadata are maintained.*

**Assessment:**

As mentioned in the description of the requirement, the operations to ensure data integrity and authenticity must be maintained throughout, starting from the ingest of data, all the way up to dissemination and access. Thereby, all three entities – DS, DIPS and ITS are responsible in fulfilling this requirement. However, each entities' roles and responsibilities towards fulfilling this requirement are not necessarily clearly defined and agreed upon by the other entities.

One of DS's responsibility is to maintain logs of provenance of the data to enable tracking of data integrity and authenticity starting at data ingest through to publication. The provenance file includes date of action, actor, and what action was taken to improve the reliability of an ingested dataset. This provenance log file is to be included in the archival bag and passed on to DIPS for preservation, but this bagging and transfer of Repository content from DS to DIPS is not currently implemented. Moreover, provenance has not been consistently captured for all published datasets (especially those first deposited).

Data Services has a version control strategy in place to ensure that any changes in published data and metadata are recorded and documented, and that published datasets are versioned appropriately**.** DS also maintains links to other datasets and metadata (implemented on Samvera) through the 'Related URLs' section and with the help of DataCite. Metadata on the repository platform is based on DublinCore. Data Services further ensures authenticity of depositors and publishers by checking their identities of through VT CAS (Central Authentication Service).

DIPS will be responsible for data integrity and running fixity checks on the data to ensure that the published and preserved digital object has not been corrupted or altered. Although, currently only the external storage services – APTrust and MetaArchive run regular fixity checks as well as maintain a log for the same. DIPS is working on incorporating a workflow for checking fixity into the preservation system. They are also developing a version control strategy and maintaining a log on all the changes made in the data and metadata. They hope to implement the two by collaborating with ITS. They are developing these functionalities on the basis of standards like the NSDA Levels of Preservation.

To ensure authenticity of the data, DIPS performs checks on preserved content in Bagit and other formats to see if they are properly structured.

ITS does not actively take part in maintaining data integrity and authenticity but maintains the systems (i.e. hardware and the Samvera application) upon which both DS and DIPS work. There is some automated tracking of changes and documentation on the Samvera platform, and checksums for files are created on ITS does not track the provenance of the data but through the Samvera platform keeps track of who deposited the data.

**Recommendations for Repository to meet/improve compliance with this Requirement:**

- The Repository should build the preservation infrastructure as planned by DIPS and integrate the Repository submission and dissemination platform within this preservation infrastructure. (DS, DIPS, ITS)
- The Repository should establish clear roles for each department regarding data integrity and authenticity, and improve coordination among all three entities. (DS, DIPS, ITS)

**Rating: 2 (The repository has a theoretical concept):**

*As detailed in the assessment, a lot of the aspects of this requirement are being implemented independently by different departments, but there is a lack of coordination among them. Many of the ideas are still under development and are yet to be implemented. (eg: fixity checks in the internal storage service).*

## Requirement 8 (Appraisal)

**Requirement Description:**

**The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.**

*The appraisal function is critical in determining whether data meet all criteria for inclusion in the collection and in establishing appropriate management for their preservation. Care must be taken to ensure that the data are relevant and understandable to the Designated Community served by the repository.*

**Assessment:**

VTechData is an institutional data repository. The Repository has minimal requirements for publishing (as given in the Publishing Requirements) that do not lead to a full appraisal of data

reusability when publishing a given dataset. The current Repository quality control checks ensure that deposited files are downloadable, that a published dataset meets all the Publishing requirements, and that deposited files can be opened (where feasible). VTechData accepts all file formats as its Designated Community is researchers of many disciplines. The administrators encourage depositors to submit their data in non-proprietary formats whenever feasible and to include metadata that allows for reusability within the expected user community, but this is not enforced strictly. Thus the checks currently done are more for the accessibility of data and not for its understandability.

As the administrators do not have the domain expertise to verify that deposited content is reusable within its originating discipline, only the above-mentioned quality checks are done.

**Recommendations for Repository to meet/improve compliance with this Requirement:**

- The Repository could leverage external consulting services such as Springer/Nature's Research Data Support or participate in the Data Curation Network project when that becomes possible. In some cases Data Services may also be able to leverage their informatics consultants for some modicum of assistance with relevance and understandability assessment. (DS)

**Rating: 0 (Not Applicable):**

*The Repository, as an Institutional Repository, is conducting the basic quality checks it can for its Designated Community (i.e. Virginia Tech researchers). However, it is beyond the scope of the Repository to mandate that published datasets are both relevant and understandable for prospective data users within the dataset's originating discipline.*

## Requirement 9 (Documented Storage Procedures)

**Requirement Description:**

**The repository applies documented processes and procedures in managing archival storage of the data.**

*Repositories need to store data and metadata from the point of deposit, through the ingest process, to the point of access. Repositories with a preservation remit must also offer 'archival storage' in OAIS terms.*

**Assessment:**

This requirement is primarily handled by ITS and DIPS. The content on VTechData is stored in two local storage locations – VTArchive (maintained by ITS) and VTCRI (Virginia Tech Carilion Research Institute) in Roanoke. Three copies of the Repository data are maintained in the Library and is handled as Library digital objects by ITS, including periodic snapshots of the Repository database. The security of the data is maintained by storing them in multiple geographical locations and the security of these storage locations is maintained at each location. ITS is not currently checking for file consistency across these copies of The Repository content.

DIPS has two external preservation storage services used for VTechData– APTrust and MetaArchive. DIPS is responsible for coordination with these two external services (DIPS Preservation Policy). All their relevant documentation on processes and handlings for storage are stored on the Confluence wiki as well as the MetaArchive and Academic Preservation Trust (APTrust) wikis.

Security of VTArchive is handled by ITS by limiting access. APTrust has a set protocol to deal with malevolent activities and prevent known threats. MetaArchive also has a Risk Analysis with protocol for preventing and handling security issues (although their documentation is not publicly accessible).

Concrete risk management techniques and methods for data recovery in the local storage systems have not been developed so far. However, APTrust and MetaArchive have well established workflows for data recovery. APTrust performs an annual test restoration of an object in our repository, and restorations can be performed at any time by preservation personnel. MetaArchive performs restorations on request. Both services conduct multiple fixity checks and hold multiple copies as part of their risk mitigation plan. VTArchive is currently implementing measures to mitigate data theft and data loss as a step towards risk management. In order to mitigate deterioration of storage media, ITS refreshes their hardware every 5 years.

While ITS maintains multiple copies of The Repository content (without fixity checking) and DIPS has a well-developed plan for creation, replication and dissemination of bagged Repository content (e.g. through APTrust and MetaArchive), Data Services has yet to pull together and bag the Repository data and metadata for preservation in the planned infrastructure. This process is ongoing.

**Recommendations for Repository to meet/improve compliance with this Requirement:**

- The Repository should replace the current hardware used, either with new hardware or through migration to cloud services. (ITS)

- The Repository should hold regular meetings to gain better insight into the operation of VTArchive and to better coordinate preservation actions. (DIPS, ITS)

**Rating: 3 (The Repository is in the Implementation Phase):**
*The Repository has well established storage locations to mitigate against accidental loss of data and has a plan for connecting published data to external preservation services. However, the Repository requires documented risk mitigation, data recovery and security protocols*

## Requirement 10 (Preservation Plan)

**Requirement Description:**

**The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.**
*The repository, data depositors, and Designated Community need to understand the level of responsibility undertaken for each deposited item in the repository. The repository must have the legal rights to undertake these responsibilities. Procedures must be documented, and their completion assured.*

**Assessment:**

The responsibility of establishing and documenting preservation procedures is mainly undertaken by DIPS in collaboration with DS. While there is a DIPS preservation services development timeline for all digital items that require archival services, including Repository content, the Repository is not yet executing necessary preservation actions. A preservation plan is in development; one of the first steps is for DIPS to internally assess the needs of VTechData and the content of the data deposited in it. As an initiation towards deciding the levels of preservation, DIPS collaborated with DS to get a clear understanding of the Preservation Profile of The Repository by looking into aspects involving content types, current preservation storage, access, search and retrieval needs.

The CoreTrustSeal requires clear understanding between the depositor (part of The Repository Designated Community) and the Repository when it comes to The Repository's responsibilities in maintaining and preserving the content. These have been clearly specified in the Publishing and Deposit Agreements given on the VTechData website. For the depositors' reference, DIPS is

developing a page on their wiki explaining the methods they will be adopting to preserve their data. The Repository can refer to this page in its Agreements. .

**Recommendations for Repository to meet/improve compliance with this Requirement:**

- The Repository should continue working together to develop a workflow for effectively bagging and preserving Repository content. (DS, DIPS)

**Rating: 3 (The Repository is in the Implementation Phase):**
*Preservation planning underway by DIPS in collaboration with DS. DIPS has already developed a preservation policy and is currently working with DS in effective preservation of Repository content.*

## Requirement 11 (Data Quality)

**Requirement Description:**

**The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.**
*Repositories must work in concert with depositors to ensure that there is enough available information about the data such that the Designated Community can assess the substantive quality of the data. Such quality assessment becomes increasingly relevant when the Designated Community is multidisciplinary, where researchers may not have the personal experience to make an evaluation of quality from the data alone. Repositories must also be able to evaluate the technical quality of data deposits in terms of the completeness and quality of the materials provided, and the quality of the metadata.*

**Assessment:**

Data Services is responsible for metadata gathering towards allowing end users to evaluate quality of published datasets. Because the Repository's Designated Community comes from a wide variety of disciplines, it is not possible for Data Services to effectively assess the reusability and quality of published datasets, as mentioned in Requirement 8.
Data and metadata that comes in through the ingest process are typically entered into or originate directly from Samvera, the Repository platform (i.e., creator, dataset/item description, file characterization and checksum). Data Services checks that the depositor has adhered to all the publishing requirements for metadata.

Deposited and published datasets include a metadata field - 'Related URLs' where depositors and publishers can include citations to related work. This is useful for depositors and publishers who will be referring to the datasets once published. DS will take feedback or complaints from Repository Users if a certain dataset is unusable for the user's purpose, and will contact the dataset creator for further information.

**Recommendations for Repository to meet/improve compliance with this Requirement:**

- The Repository could convene a team of disciplinary experts to help assess the quality of the content in the dataset, and leverage external efforts to develop guidance for assessing dataset fitness for use (e.g. the work of the WDS/RDA Assessment of Data Fitness for Use WG). (DS, See Requirement 8)

**Rating: 0 (Not Applicable):**

*As The Repository is an institutional repository, it hosts data from varied disciplines and thus it is not currently practical to assess the quality of published datasets.*

## Requirement 12 (Workflows)

**Requirement Description:**

***Archiving takes place according to defined workflows from ingest to dissemination.***

*To ensure the consistency of practices across datasets and services and to avoid ad hoc and reactive activities, archival workflows should be documented, and provisions for managed change should be in place. The procedure should be adapted to the repository mission and activities, and procedural documentation for archiving data should be clear.*

**Assessment:**

We can separate the curation process into two parts: dataset ingest and publication workflow and dataset preservation workflow. The dataset ingest and publication workflow is managed by DS while the dataset preservation workflow is managed by DIPS.

DS has an established workflow which encompasses ingest, curation, publishing, and a portion of the preservation process. This workflow document, while complete, needs more detail about each curation step. DS ensures that the files are openable and accessible and adhere to the Deposit Agreement before proceeding to publish the data. Many stages in the current workflow are done manually by DS and the dataset depositor. The workflow is the same for all datasets with files smaller than 2GB. If the size of files exceed this 2GB limit, Data Services curates the dataset (or a portion thereof) using Globus Services.

DIPS is developing the preservation workflow and are building a workflow common to all digital platforms at Virginia Tech including The Repository. The documentation, which is still under progress can be found on their Wiki.

**Recommendations for Repository to meet/improve compliance with this Requirement:**

- The Repository should improve detail of documentation for all workflows, and ideally such that new personnel could execute these workflows without further guidance. (DS, DIPS, ITS)
- The Repository should automate the portions of the ingest, publication and preservation workflow that can be automated. (DS, DIPS, ITS)

**Rating: 2 (The repository has a theoretical concept):**

*Although the first half of the Repository workflow (ingest and publishing) has been established and is being refined, the preservation workflow is at the beginnings of implementation. There is also more room for improvement with respect to the ingest and publishing workflow as many stages are yet to be automated and fully documented.*

## Requirement 13 (Data discovery and identification)

**Requirement Description:**

**The repository enables users to discover the data and refer to them in a persistent way through proper citation.**

*Effective data discovery is key to data sharing, and most repositories provide searchable catalogues describing their holdings such that potential users can evaluate data to see if they meet their needs. Once discovered, datasets should be referenceable through full citations to the data, including persistent identifiers to ensure that data can be accessed into the future. Citations also provide credit and attribution to individuals who contributed to the creation of the dataset.*

**Assessment:**

ITS has worked with DS to provide dataset citations on The Repository platform that include persistent identifiers (DOIs obtained through DataCite), and allow credit and attribution to individuals as part of data discovery and identification (see an example here). In order to

facilitate search options, the repository has an internal search box on Samvera that searches some (but not all) metadata fields. There is currently no API that enables the metadata to be searched externally, but datasets are discoverable through DataCite and the DataCite API. There is also a metadata search catalogue that is maintained in the Samvera application and it is accessible through the "Search" toolbar on the Repository website. VTechData is registered in re3data.org, the registry for research data repositories.

**Recommendations for Repository to meet/improve compliance with this Requirement:**

- The Repository should allow metadata to be machine accessible directly through the Repository Platform. (DS, ITS)

**Rating: 4 (The guideline has been fully implemented in the repository):**
*This requirement seems to be fully implemented as it has met most of the specifications given by CoreTrustSeal.*

## Requirement 14 (Data Reuse)

**Requirement Description:**

**The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.**
*Repositories must ensure that data can be understood and used effectively into the future despite changes in technology. This Requirement evaluates the measures taken to ensure that data are reusable.*

*From the CoreTrustSeal Extended Guidance: The applicant should understand the needs of the Designated Community in terms of their research practises, technical environment, and applicable standards.*

**Assessment:**

The responsibility of DS when it comes to Repository data re-use is to encourage depositors to provide metadata for their intended audience while sticking to appropriate community standards where possible. Metadata fields are adopted from Dublin Core as it is used by the Samvera platform.

VTechData accepts all file formats as its Designated Community is researchers of many disciplines. Therefore, there are no specific sets of data formats or metadata standards specified for published datasets. As there is no way for the Repository to assess the understandability of the datasets due to the varied disciplines and formats they come in, DS encourages the depositors to provide appropriate metadata for those with which they intend to share their data. While DS can provide general guidance for data management across the disciplines at Virginia Tech, DS cannot be an expert in all disciplines' research practises, technical environment, and applicable standards.

DIPS provides or intends to enable access to published data and metadata over time. DIPS does not generate metadata about each Repository dataset as that is a DS responsibility, but DIPS does generate preservation metadata (i.e. metadata generated and stewarded to enable Libraries preservation actions now and in the future).

When a Repository dataset is to be preserved, DIPS creates a UUID and obtains the Local ID from the content manager for the metadata. The preservation metadata is read into CSV or XML formats. In a scenario where there is a possibility for evolution of the formats, DIPS would like to automate format migration.

**Recommendations for Repository to meet/improve compliance with this Requirement:**

- The Repository could continue developing a preservation metadata server and create/convert all preservation metadata into RDF. (DIPS, ITS)

**Rating: 0 (Not Applicable)**

*As The Repository receives data from all across the University (and from multiple disciplines), DS is not in a position to ensure that the data is reusable either now or in the future. DS ensures whatever data and metadata they received during ingest and publishing is accessible in the future.*

## Requirement 15 (Technical infrastructure)

**Requirement Description:**

**The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.**

*Repositories need to operate on reliable and stable core infrastructures that maximizes service availability. Furthermore, hardware and software used must be relevant and appropriate to the*

*Designated Community and to the functions that a repository fulfils. Standards such as the OAIS reference model specify the functions of a repository in meeting user needs.*

**Assessment:**

ITS is responsible for providing storage, maintaining storage and server infrastructure, and planning a sustainable infrastructure meeting the Repository needs. The Repository standards theoretically conform to OAIS (Open Archival Information System) though further development is needed.

All software currently in use for the Repository production instance are maintained in two GitHub repos – one for VTechData and one for InstallScripts, both of which are publicly available. The majority of the Repository platform (currently Samvera) is largely based on open source software, with exceptions of certain software libraries.

According to ITS, the current physical Repository infrastructure is at the tail-end of its active development. They are currently looking into incorporating new components into the infrastructure which is expected to be a joint effort between DS, ITS and DIPS and are formalizing plans for this effort., Although statistics are not being retained, the up-time for the Repository is monitored constantly and is considered satisfactory.

One major issue with the current technical infrastructure is that it is heavily reliant on Samvera and is built on a Fedora database, which according to ITS is not up to the mark. Data Services states that the current Samvera interface is not satisfactory for usage by the Designated Community for deposit and publication. The Repository software is heavily customized and is difficult to update.

**Recommendations for Repository to meet/improve compliance with this Requirement:**

- The Repository should complete documentation of its technical infrastructure. (ITS)
- The Repository should move away from usage of the Samvera platform for ingest and dissemination of content. (DS, ITS)

**Rating: 3 (The repository is in the implementation phase)**

*Most of the infrastructure development and maintenance is fully implemented but not completely documented. ITS and DS are currently investigating other Repository platforms to use in the future.*

## Requirement 16 (Security)

**Requirement Description:**

**The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.**
*The repository should analyze potential threats, assess risks, and create a consistent security system. It should describe damage scenarios based on malicious actions, human error, or technical failure that pose a threat to the repository and its data, products, services, and users. It should measure the likelihood and impact of such scenarios, decide which risk levels are acceptable, and determine which measures should be taken to counter the threats to the repository and its Designated Community. This should be an ongoing process.*

**Assessment:**
ITS takes the responsibility of updating hardware and software periodically and conducting threat analysis. Threat analysis is mostly done as an ad-hoc process and includes virus protection, and protecting against Repository administrative errors (e.g. Data Services deleting published datasets).

In case of an outage, ITS has procedures in place to bring up the service on a secondary site with a 24-hour old copy of the data. Essentially, data that had been uploaded within 24 hours will only be lost and the rest will be retained. ITS has a security system which has an in-depth approach with multiple layers of security. As part of the disaster recovery, multiple copies of the data are maintained (CRC and Library).

In order to keep the data secure, DIPS ensures that the copies are maintained at different geographical locations. They also provide for data recovery using checksum discrepancies. DIPS has multiple external entities (e.g. Archivematica, MetaArchive, see Requirement 9) who are responsible for disaster recovery, business continuity planning, and threat analysis for Repository content they steward. The Repository's content is not yet incorporated into DIPS preservation workflows and thus not yet connected with these external entities.

**Recommendations for Repository to meet/improve compliance with this Requirement:**

- The Repository should complete documentation of its technical infrastructure. (ITS)
- The Repository should should formalize its succession plan and data preservation plans and workflows. (DS, DIPS, ITS)
- The Repository should clarify responsibilities and roles regarding security. (DIPS, ITS)

**Rating: 3 (The repository is in implementation phase)**

*The repository has a disaster recovery and other security measures in place but lacks good documentation. This is something that can be implemented soon.*