The Institutional Repository's Role in Preserving Research Data

Zhiwu Xie, Gail McMillan, and Tyler Walters

VirginiaTech | University Libraries

Introduction

Many funding agencies have started to require long-term preservation and open access to research data. While most research universities have already run their own institutional repositories (IR), it's not clear what role the IR can play in preserving these data.

Unlike the textual and even multimedia contents currently archived by the conventional IR, research data are much more diverse in terms of format, metadata, storage, rendering, and access requirements. The differences between the geospatial data, astronomical observation data, DNA sequencing data, and computational fluid dynamics simulation data can be so large as to deserve their own disciplinary data repositories. A disciplinary repository can customize its structure and functionality for a specific type of data, a luxury not available to the general-purpose IR.

On the other hand, the IR is uniquely positioned to preserve and manage the research data. The institute provides the IT infrastructure where most of these data are initially generated, processed, stored, and managed. As part of the IT infrastructure, the IR usually presents the lowest entrance and migration barrier as well as the cheapest cost for data created within the same institution.
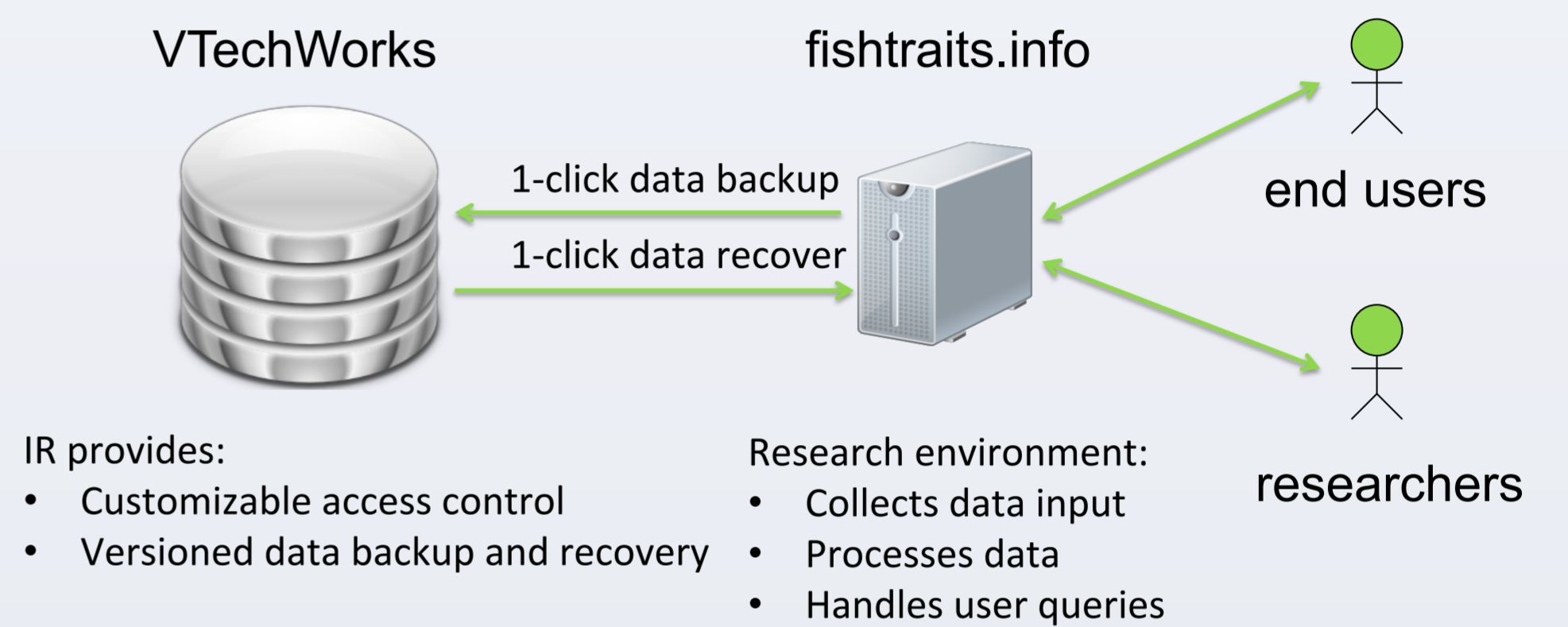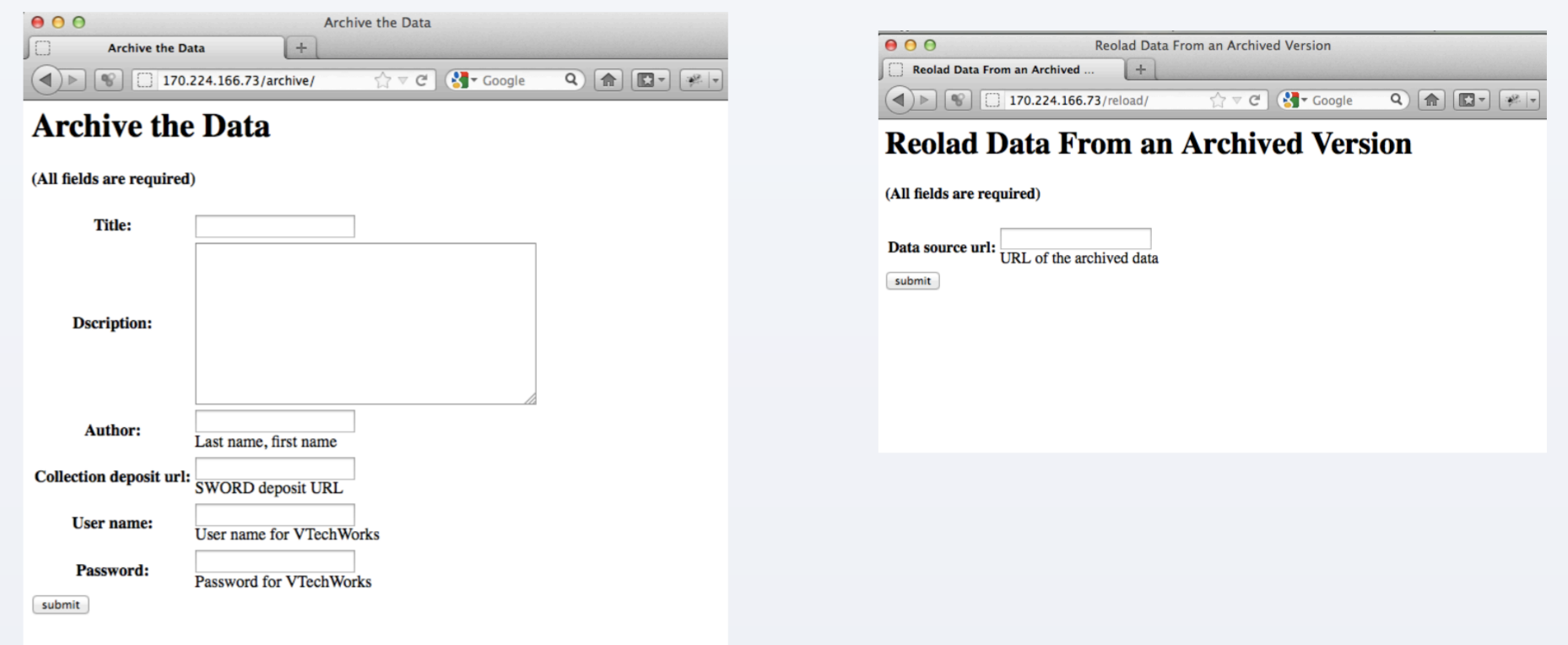
The IR's Role

In order to meet the data managing challenges, we must clearly define the core functionality an IR must provide during the lifecycle of the research data, which may include:
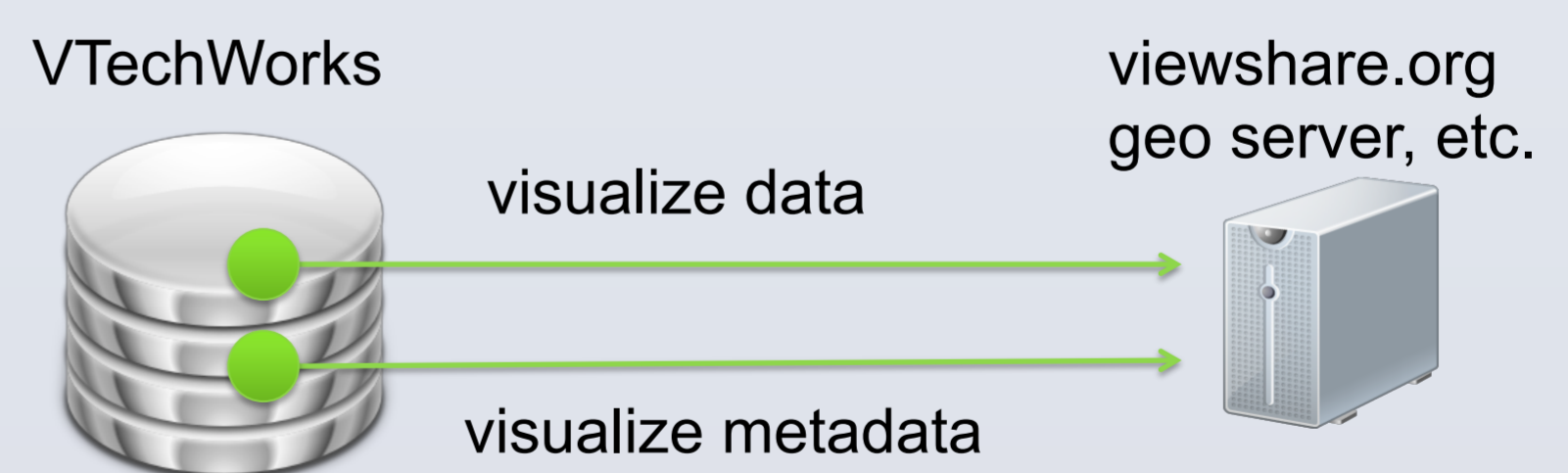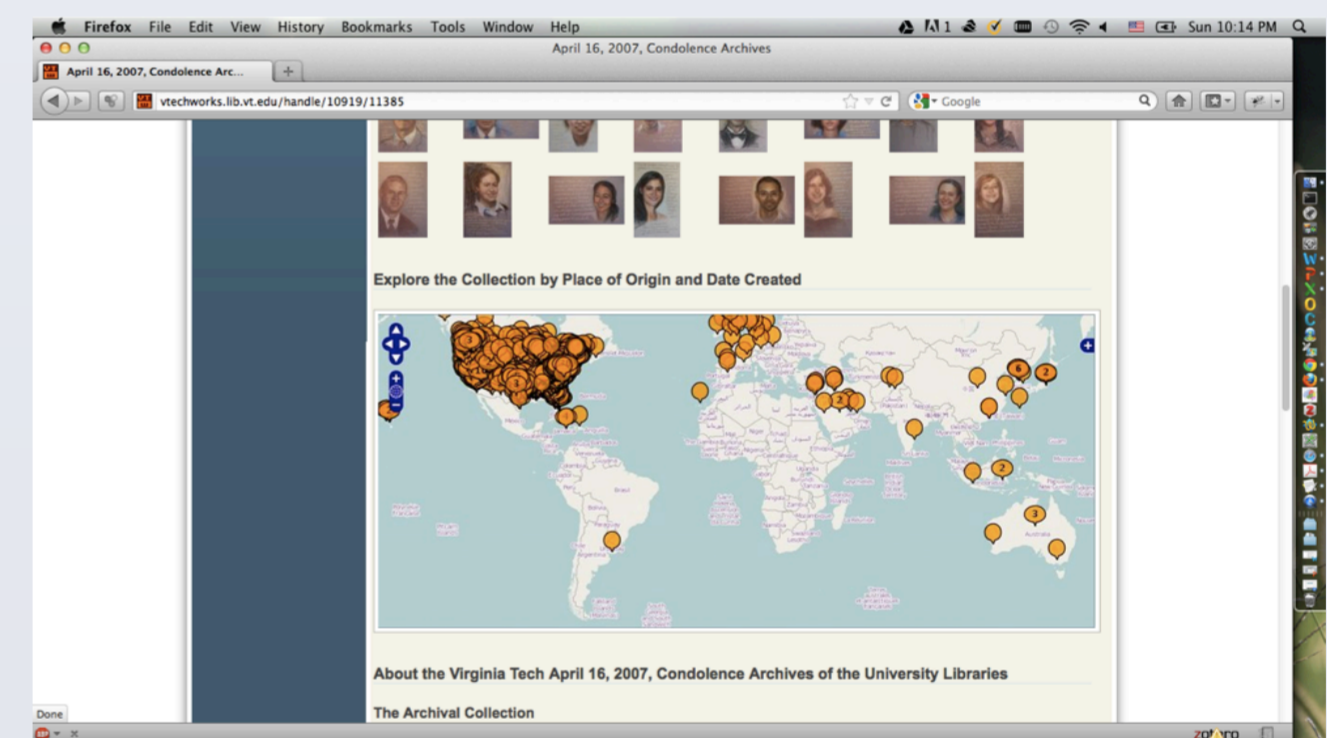
- Closely integrate the IR with the university's IT infrastructure to allow easy deposit and access control
- Provide the baseline storage needs, which may be further differentiated by the usage pattern to lower the cost
- Act as a metadata hub that not only can understand various disciplinary metadata, but can also translate them into more widely understood terms for easy discovery and access
- Facilitate reuse and preservation by at least maintaining the preservation metadata that document the environment where the data originally lived
- Provide programming interfaces to facilitate the data visualization, presentation, and usage from external services
- Provide data exchange interfaces to various disciplinary data repositories

Virginia Tech is working towards building its IR, VTechWorks, as an exemplary general-purpose repository that fulfills these data management roles.

Example: Tighter Integration with the Research Environment

VTechWorks   fishtraits.info   end users

1-click data backup
1-click data recover

researchers

IR provides:
- Customizable access control
- Versioned data backup and recovery

Research environment:
- Collects data input
- Processes data
- Handles user queries

Example: Visualize the Data Using External Services

VTechWorks   viewshare.org geo server, etc.

visualize data

visualize metadata

Example: Data Exchange

VTechWorks   DataOne, BioMed Central, etc.

exchange data