

SDB: A Cross Industry, Flexible and Scalable Preservation Architecture

Mark Evans, Tessella Inc

Oracle PASIG, May 11th 2011



Contents

- Tessella archiving history
- Cross Industry Considerations
- SDB4 Architecture:
 - Flexibility
 - Scalability
- Current research programs
- Conclusions

Brief History

- Tessella have been working in digital archiving for over a decade:
 - Mostly with memory institutions
 - Recent engagements with life sciences and scientific research
- Out of this effort Safety Deposit Box (SDB) has grown:
 - 12 customers
 - Utilized output from Planets
 - Now on version 4
 - Product roadmap
 - Support team
 - SDB Users Group

SDB4 Solutions Worldwide



The National Archives

UK National Archives



Wellcome Trust



Finnish National Archives



Estonian National Archives



Gemeente Rotterdam
Gemeentearchief

Rotterdam City Archive

nationaal archief

Dutch National Archives



Malaysian Archives



FamilySearch



STFC



Schweizerisches Bundesarchiv
Swiss Federal Archives

FEDERAL CHANCELLERY AUSTRIA
Austrian Government



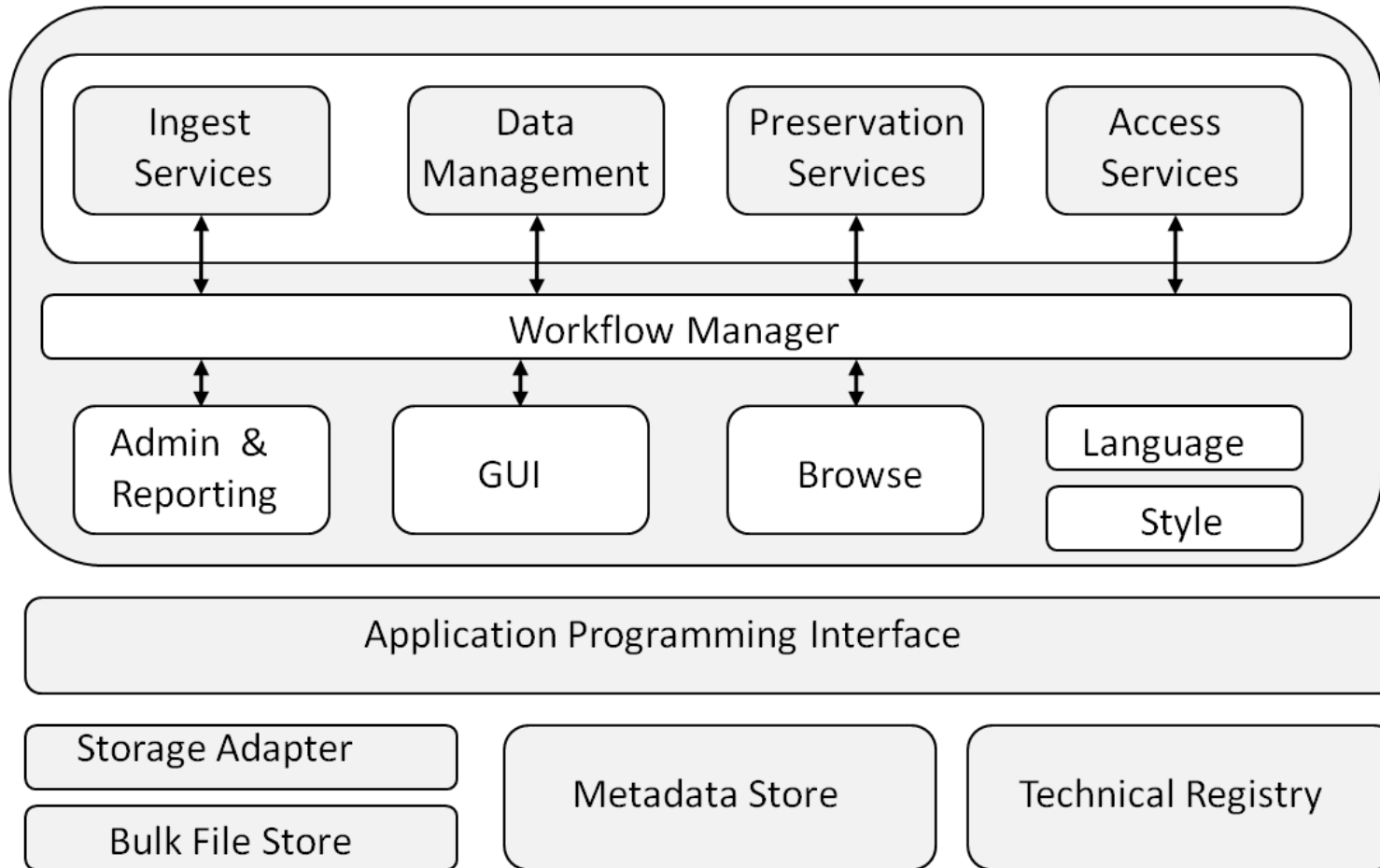
Why Do Digital Archiving

- Avoid damage to organisation:
 - Need to comply with regulatory requirements
 - Defend legal claims / patent infringements etc.
 - Reputation: Need to be seen to treat information with respect
 - Cost of maintaining existing systems prohibitive
- Gain benefits:
 - Need to reuse information, support eDiscovery
- Applies to everyone but in particular:
 - Pharmaceutical
 - Health care
 - Financial
 - Aerospace
 - Nuclear
 - Oil/gas

Demands of Other Domains

- Everything in archives / libraries etc
 - All of OAIS
 - Records Management functions
- Flexibility:
 - Take content from different sources in many different formats
 - Structured (data) as well as unstructured (documents)
 - Often in highly specialised formats / custom databases etc.
 - Privacy very important
- Scalability:
 - Hundreds of thousands of employees:
 - Process huge volumes, preferably at short notice
- Need cost/benefit analysis

Safety Deposit Box



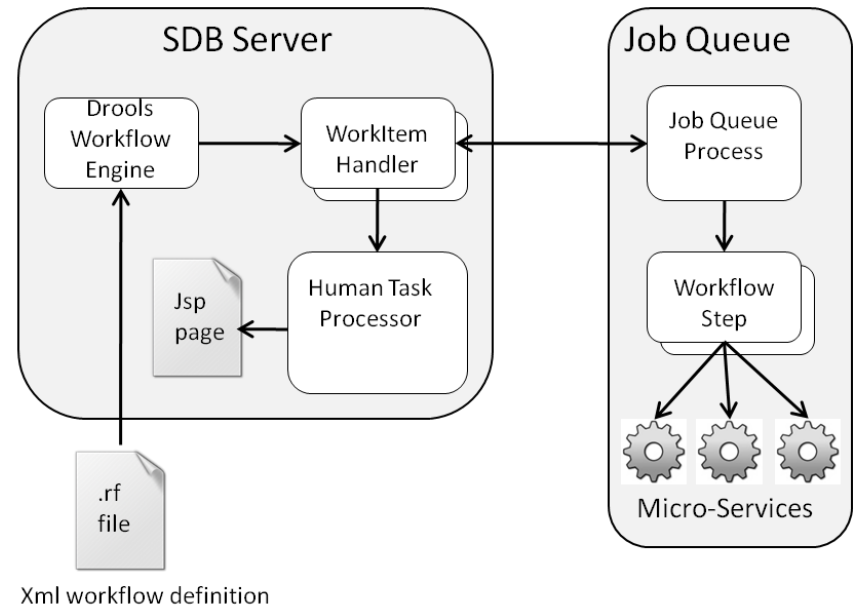
Workflow

- Linear workflow tool:

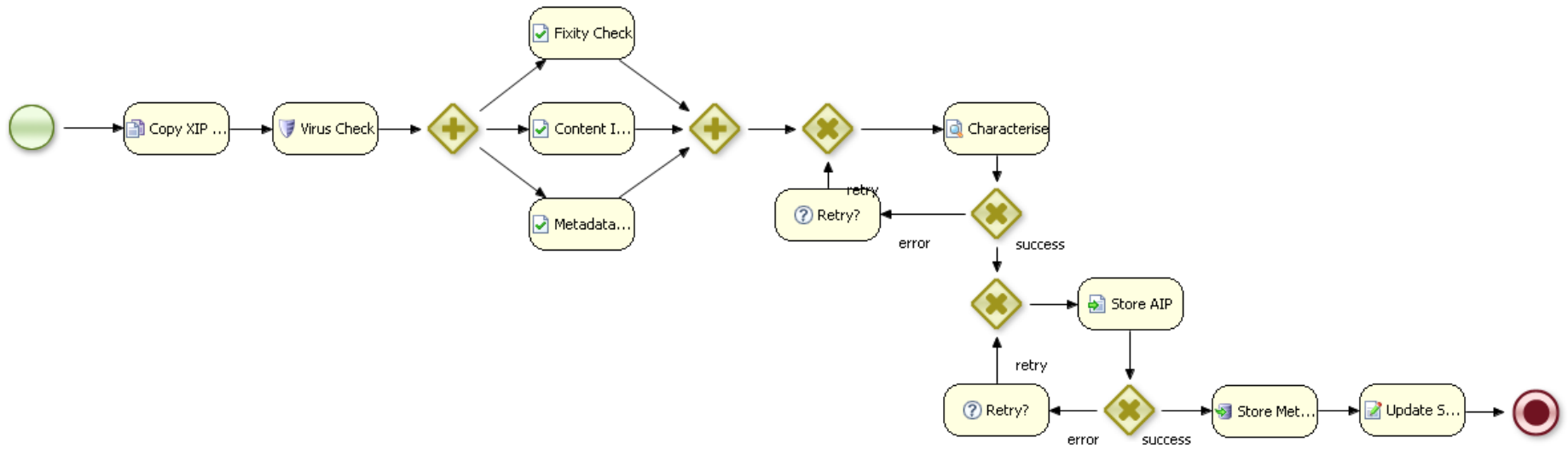
- Create workflows
- Schedule workflows
- Alter criteria (parameters)
- Handle errors
- 'Acceptable errors'

- SDB4 Workflow engine

- Drools open source rules engine
- Includes Splits/Joins
- Wait States/Timers
- Human tasks
- Workflow can be generated by "Drag and Drop"



Example Ingest workflow



Workflow Management


Firefox | Tessella SDB | http://sdbpilot.tessella.com/sdb/ingest.html#

Most Visited | Getting Started | Latest Headlines | QFORG_01.htm | Location, Hours and D... | Bookmarks

Tennessee State Library & Archives | Digital Preservation
National Digital Information Infrastructure & Preservation Program
A Collaborative Initiative of the Library of Congress

Welcome, Mark Evans TN (Tenant : TENNESSEE) Logout | SDB Digital Archive: Ingest | Home

Start | **Waiting** | Running | Completed | Reports | Manage

 Filter Workflows

Submission name	Collection Code	Top Level Record	Date Created ↓	Agency	Size	Files	Workflow Context
ZIP Upload			28.04.11 13:31:11			0	ZIP Upload
ZIP Upload			28.04.11 13:22:20			0	ZIP Upload
ZIP Upload			28.04.11 13:17:31			0	ZIP Upload
ZIP Upload			28.04.11 13:15:30			0	ZIP Upload

Help | tessella.com | Copyright © 2010 Tessella



Workflow Progress

Firefox | Tessella SDB | http://sdbpilot.tessella.com/sdb/workflowProgress.html?wkid=7880

Most Visited | Getting Started | Latest Headlines | QFORG_01.htm | Location, Hours and D... | Bookmarks

Start | **Waiting** | **Running** | **Completed** | **Reports** | **Manage**

Workflow Details

Workflow Context	ZIP Upload
Workflow Definition	Ingest Workflow (Manual Upload, No Virus Scan)
Workflow ID	7880
Workflow State	Aborted
Date Started	28.04.11 18:40:23
Date Finished	02.05.11 23:01:48
Number of Files	2
Total Size	424 KB
Collection Code	CRK VT
Submission name	CRK VT
Top Level Record	New Folder

Step Progress

State	Name	Progress	Started	Finished	Messages
	Upload SIP Package	<div style="width: 100%; height: 10px; background-color: green;"></div>	28.04.11 18:40:23	02.05.11 23:01:38	
	Unzip SIP	<div style="width: 100%; height: 10px; background-color: green;"></div>	02.05.11 23:01:38	02.05.11 23:01:40	
	Fixity Check	<div style="width: 100%; height: 10px; background-color: green;"></div>	02.05.11 23:01:40	02.05.11 23:01:42	
	Metadata Integrity	<div style="width: 100%; height: 10px; background-color: green;"></div>	02.05.11 23:01:42	02.05.11 23:01:44	
	Content Integrity	<div style="width: 100%; height: 10px; background-color: green;"></div>	02.05.11 23:01:44	02.05.11 23:01:46	
	Characterise	<div style="width: 100%; height: 10px; background-color: red;"></div>	02.05.11 23:01:46	02.05.11 23:01:48	View

6:19 PM 5/11/2011

SDB Flexible Deployment Options

- Local deployment
 - SDB, Storage, Database
- Hosted solution
 - Tessella hosted, 3rd party host
- Cloud based solution
 - Storage – S3,etc (Now)
 - SaaS (future)
- Multi Tenancy
 - Shared instance
 - Each tenant can control content, policy, functionality etc

SDB4: Cross Industry Flexibility

- Choose ingest source:
 - EDRMS / Content management
 - Workflow systems
 - Web sites
 - Flat files & catalogue
- Choose descriptive metadata schema:
 - DON'T convert
 - Support for any schema
 - Still allow view / edit / fielded search
 - Plus synchronisation with external catalogues (e.g., via OAI-PMH)

Cross-Industry Flexibility

- Choose functionality (via workflow system):
 - Can add new steps
 - Can create new workflows
- Choose configuration / security (multiple tenancy):
 - Single administered instances
 - Multiple organisations / departments
- Choose storage system and AIP structure:
 - Use existing or add new storage adaptor
- Choose database engine:
 - Oracle, mySQL, SQL Server, ...
- Choose reporting options

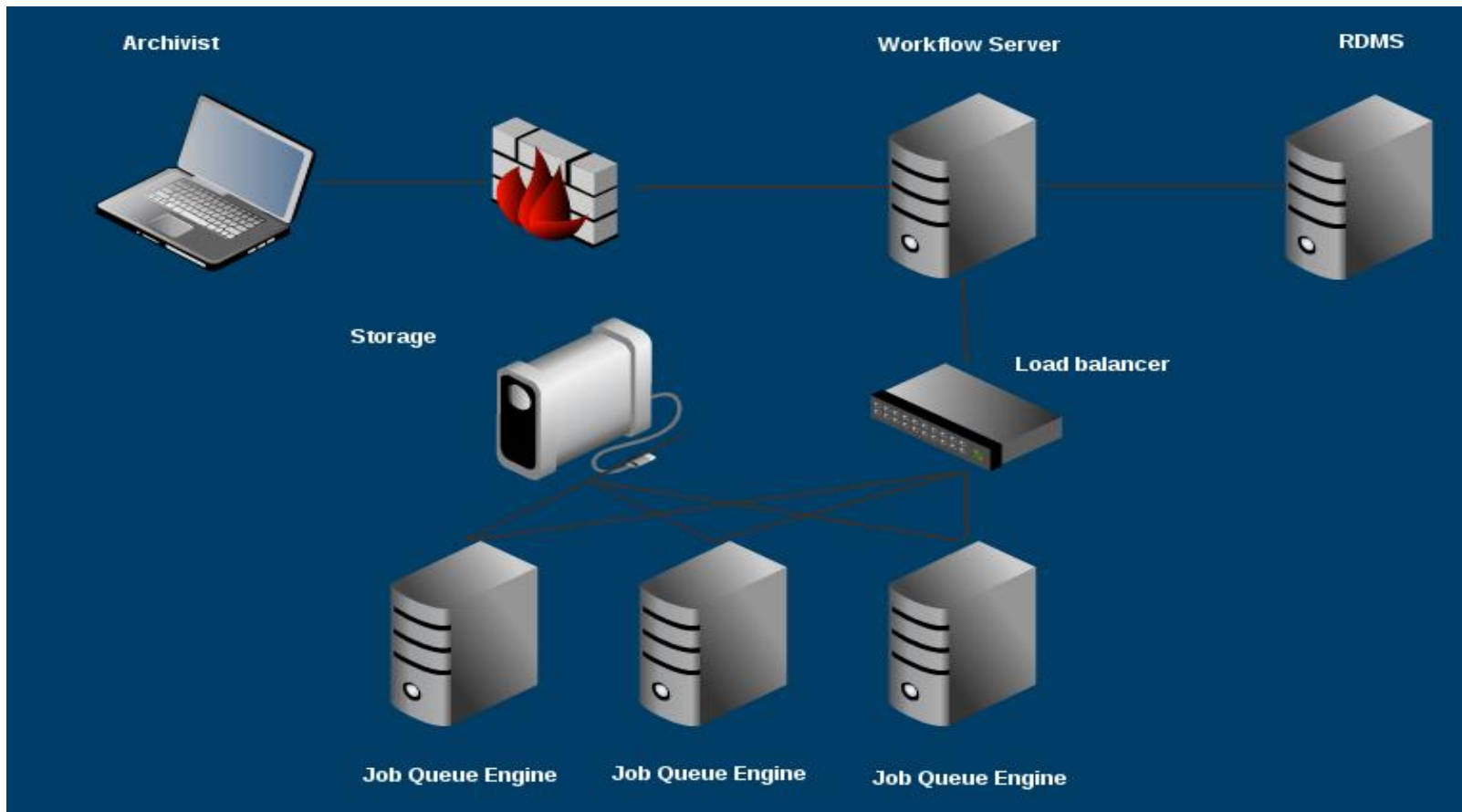
SDB4: Cross Industry Flexibility: Example

- Choose characterisation functionality:
 - Format identification tool
 - Format validation tool (per format)
 - Property extraction tool (per format)
 - Embedded object extraction tool (per format)
 - Logical characterisation tool
- Choose preservation functionality:
 - What's at risk?
 - Migration pathway / tool
 - Validation criteria

SDB4: Cross Industry Scalability

- Lots of long-running jobs:
 - Farm out to multiple servers
 - Utilise job queuing system (control threads per server)
- How fast can we ingest?:
 - Used Oracle hardware and database in a test suite hosted by Oracle
 - Test data:
 - Thousands of 1GB SIPs (100 c.10MB files each)
 - Mix of formats (PDF, TIFF, JPEG)
 - Workflow:
 - Copy from source
 - Fixity check
 - Integrity checks
 - Characterise
 - Store content
 - Store metadata

SDB4: Cross Industry Scalability



SDB4: Cross Industry Scalability

- Tuned system parameters:
 - Built performance model
- Achieved 2TB/day per server (SunFire X4140):
 - BUT local server almost idle.
 - Held up by speed of reading content from source
 - Network also close to saturation
 - Hence, adding more job queue servers didn't help

SDB4: Cross Industry Scalability

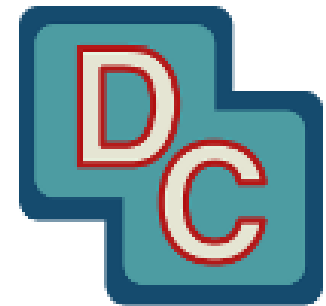
- Working with FamilySearch:
 - 4.4GB test SIPs (c. 10MB JPEG2000 files)
 - Similar workflow
 - Need c. 20 TB/day
- Initially similar barrier, so:
 - Updated storage array (ingest queue) to 168 disks in parallel
 - Don't move content more than needed
 - Added second job queue server
- Now achieved:
 - > 20 TB/day
- Tuning work will continue



Current Research Activities

- Ensure:
 - Started 1-FEB-2011, 3 years
- Apply digital preservation to
 - Health Care
 - Clinical Trials
 - Financial Data
- Aparsen:
 - Develop a network of excellence / Best Practice
- DataNet – Data Conservancy / DataOne
 - Sustainable curation and preservation cyberinfrastructure
 - Focus on Scientific Data

ENSURE



Conclusions

- Cross-organisation needs vary within memory institutions
- Cross-industry demands are also varied
- Generally all domains demand:
 - Flexibility
 - Scalability
- SDB4 is designed to meet these demands
- Research underway to demonstrate digital preservation in “new” domains