HOW TO STORE YOUR DATA TO ENABLE MODERN DATA ANALYTICS: NASA'S HARDWARE INSPECTION DATA

QUALITY LEADERSHIP FORUM

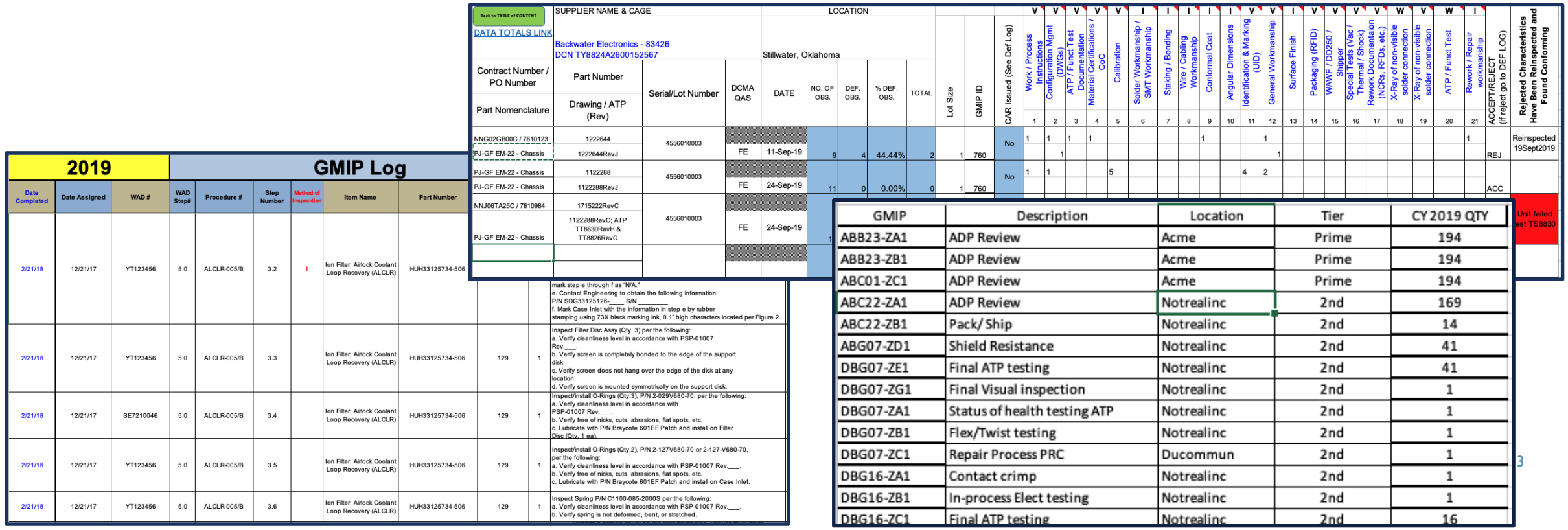
2021-05-01

By Elizabeth Wainwright & Justin Gosses
NASA OCIO Transformation and Data Division Data Analytics

CONTENTS

CONTEXT: INSPECTION DATA HISTORICALLY

- Each hardware inspection report is generated individually. Differences between reports are significant.
- No analysis is typically done across large numbers of reports.



2019 GMIP Log

Date Completed	Date Assigned	WAD #	WAD Step#	Procedure #	Step Number	Method of Inspection	Item Name	Part Number			
2/21/18	12/21/17	YT123456	5.0	ALCLR-005/B	3.2	I	Ion Filter, Airlock Coolant Loop Recovery (ALCLR)	HUH33125734-506			mark step e through f as "N/A." e. Contact Engineering to obtain the following information: P/N SDG3312512 6-____ S/N _____ f. Mark Case Inlet with the information in step e by rubber stamping using 73X black marking ink, 0.1" high characters located per Figure 2.
2/21/18	12/21/17	YT123456	5.0	ALCLR-005/B	3.3		Ion Filter, Airlock Coolant Loop Recovery (ALCLR)	HUH33125734-506	129	1	Inspect Filter Disc Assy (Qty. 3) per the following: a. Verify cleanliness level in accordance with PSP-01007 Rev.___. b. Verify screen is completely bonded to the edge of the support disk. c. Verify screen does not hang over the edge of the disk at any location. d. Verify screen is mounted symmetrically on the support disk.
2/21/18	12/21/17	SE7210046	5.0	ALCLR-005/B	3.4		Ion Filter, Airlock Coolant Loop Recovery (ALCLR)	HUH33125734-506	129	1	Inspect/install O-Rings (Qty.3), P/N 2-029V680-70, per the following: a. Verify cleanliness level in accordance with PSP-01007 Rev.___. b. Verify free of nicks, cuts, abrasions, flat spots, etc. c. Lubricate with P/N Braycote 601EF Patch and install on Filter Disc (Qty. 1 ea).
2/21/18	12/21/17	YT123456	5.0	ALCLR-005/B	3.5		Ion Filter, Airlock Coolant Loop Recovery (ALCLR)	HUH33125734-506	129	1	Inspect/install O-Rings (Qty.2), P/N 2-127V680-70 or 2-127-V680-70, per the following: a. Verify cleanliness level in accordance with PSP-01007 Rev.___. b. Verify free of nicks, cuts, abrasions, flat spots, etc. c. Lubricate with P/N Braycote 601EF Patch and install on Case Inlet.
2/21/18	12/21/17	YT123456	5.0	ALCLR-005/B	3.6		Ion Filter, Airlock Coolant Loop Recovery (ALCLR)	HUH33125734-506	129	1	Inspect Spring P/N C1100-085-2000S per the following: a. Verify cleanliness level in accordance with PSP-01007 Rev.___. b. Verify spring is not deformed, bent, or stretched.

Back to TABLE of CONTENT

DATA TOTALS LINK

SUPPLIER NAME & CAGE: Backwater Electronics - 83426 DCN TY8824A2600152567

LOCATION: Stillwater, Oklahoma

Contract Number / PO Number / Part Nomenclature	Part Number / Drawing / ATP (Rev)	Serial/Lot Number	DCMA QAS	DATE	NO. OF OBS.	DEF. OBS.	% DEF. OBS.	TOTAL	Lot Size	GMIP ID	CAR Issued (See Def Log)	V 1 Work / Process Instructions	V 2 Configuration Mgmt (DWGs)	V 3 ATP / Funct Test Documentation	V 4 Material Certifications / CoC	V 5 Calibration	I 6 Solder Workmanship / SMT Workmanship	I 7 Staking / Bonding	I 8 Wire / Cabling Workmanship	I 9 Conformal Coat	I 10 Angular Dimensions	V 11 Identification & Marking (UID)	V 12 General Workmanship	I 13 Surface Finish	V 14 Packaging (RFID)	V 15 WAWF / DD250 / Shipper	V 16 Special Tests (Vac / Thermal / Shock)	V 17 Rework Documentation (NCRs, RFDs, etc.)	W 18 X-Ray of non-visible solder connection	V 19 X-Ray of non-visible solder connection	W 20 ATP / Funct Test	I 21 Rework / Repair workmanship	ACCEPT/REJECT (If reject go to DEF LOG)	Rejected Characteristics Have Been Reinspected and Found Conforming
NNG02GB00C / 7810123	1222644	4556010003																																
PJ-GF EM-22 - Chassis	1222644RevJ		FE	11-Sep-19	9	4	44.44%	2	1	760	No	1	1	1	1					1			1								1	REJ	Reinspected 19Sept2019
PJ-GF EM-22 - Chassis	1122288	4556010003																																
PJ-GF EM-22 - Chassis	1122288RevJ		FE	24-Sep-19	11	0	0.00%	0	1	760	No	1	1			5						4	2									ACC	
NNJ06TA25C / 7810984	1715222RevC	4556010003																																
PJ-GF EM-22 - Chassis	1122288RevC; ATP TT8830RevH & TT8826RevC		FE	24-Sep-19	1																													Unit failed est TS8830

GMIP	Description	Location	Tier	CY 2019 QTY
ABB23-ZA1	ADP Review	Acme	Prime	194
ABB23-ZB1	ADP Review	Acme	Prime	194
ABC01-ZC1	ADP Review	Acme	Prime	194
ABC22-ZA1	ADP Review	Notrealinc	2nd	169
ABC22-ZB1	Pack/ Ship	Notrealinc	2nd	14
ABG07-ZD1	Shield Resistance	Notrealinc	2nd	41
DBG07-ZE1	Final ATP testing	Notrealinc	2nd	41
DBG07-ZG1	Final Visual inspection	Notrealinc	2nd	1
DBG07-ZA1	Status of health testing ATP	Notrealinc	2nd	1
DBG07-ZB1	Flex/Twist testing	Notrealinc	2nd	1
DBG07-ZC1	Repair Process PRC	Ducommun	2nd	1
DBG16-ZA1	Contact crimp	Notrealinc	2nd	1
DBG16-ZB1	In-process Elect testing	Notrealinc	2nd	1
DBG16-ZC1	Final ATP testing	Notrealinc	2nd	16

This is faked data but uses real excel layouts

CONTEXT:
Problems With 96 Historical Inspection Data Files from 2019

How data is represented

- **File Type Inconsistency**: 90% excel, 10% PDF.
- **Sheet Number Inconsistency**: Varying numbers of sheets, and no definitions for them.
- **Tables , Forms**: Large variance in structure of how data is recorded between almost every inspection instance. Few if any are designed to be machine readable.
- **Multiple Headers**: Inconsistent numbers of headers make data difficult to parse
- **Non-text Information**: Color and images prevent are not machine readable
- **Inconsistent Column Names**: Different names for the same field & Different meaning for the same field. Almost nothing defined their terms.

What data is captured

- Different information (fields) are captured in each inspection report!

CONTEXT:
QUESTIONS WE WANT TO BE ABLE TO ASK IN FUTURE!

Types of questions that require the ability to analysis across many inspection reports

Average time to filing of completed inspection report from inspection request?

Inspection time to complete correlates with what?

What company does the most inspections?

How much do inspection rates vary across systems, type of hardware, or vendor?

How doe failure rates vary by inspection type?

What inspection type has seen the biggest improvement in failure rates?

What controls are more or less variable within the same inspection process type ?

CONTEXT:
DATA ANALYTICS PREVENTED BY POOR DATA MANAGEMENT

Whether final solution is Excel template, web application, or something else, the main problems to solved are:

- *Maximize number of questions that can be answered across inspection reports through standardization of fields to the greatest extent practical.*
- *Build a solution that takes into account the real irreducible variation in data captured by inspection reports such that information is both captured and doesn't result in poor data quality or inability to ingest data programmatically.*

CONTEXT: OUR TASK

- **Problem:** There is need to be able to do data analytics across a large number of inspection reports, but every hardware inspection report is different. As a result, it is impossible to do analysis on them in aggregate.

- **Goals:**
  1. **Understand** the current state of hardware inspection data.
  2. **Recommend** data management processes and technology to enable data analytics
  3. **Build** some prototypes to explore the solution space

- **Constraints:**
  1. Only two people will be working on this project.
  2. Initially limited assistance from people with domain knowledge.
  3. Our part won't by itself produce a final product / workflow but rather to identify the characteristics needed to handle the data variance & enable modern data analytics.

CONTEXT

Where We Are Now

1) An analysis of the current GMIP data
   - Completed
2) A initial proposal for data standards for GMIP data
   - 90% done, needs technical work & review/agreement with procurement

Going Forward

1) Integration of this work with Goddard Meta team that will be building SCIS (Supply Chain Insight Central)

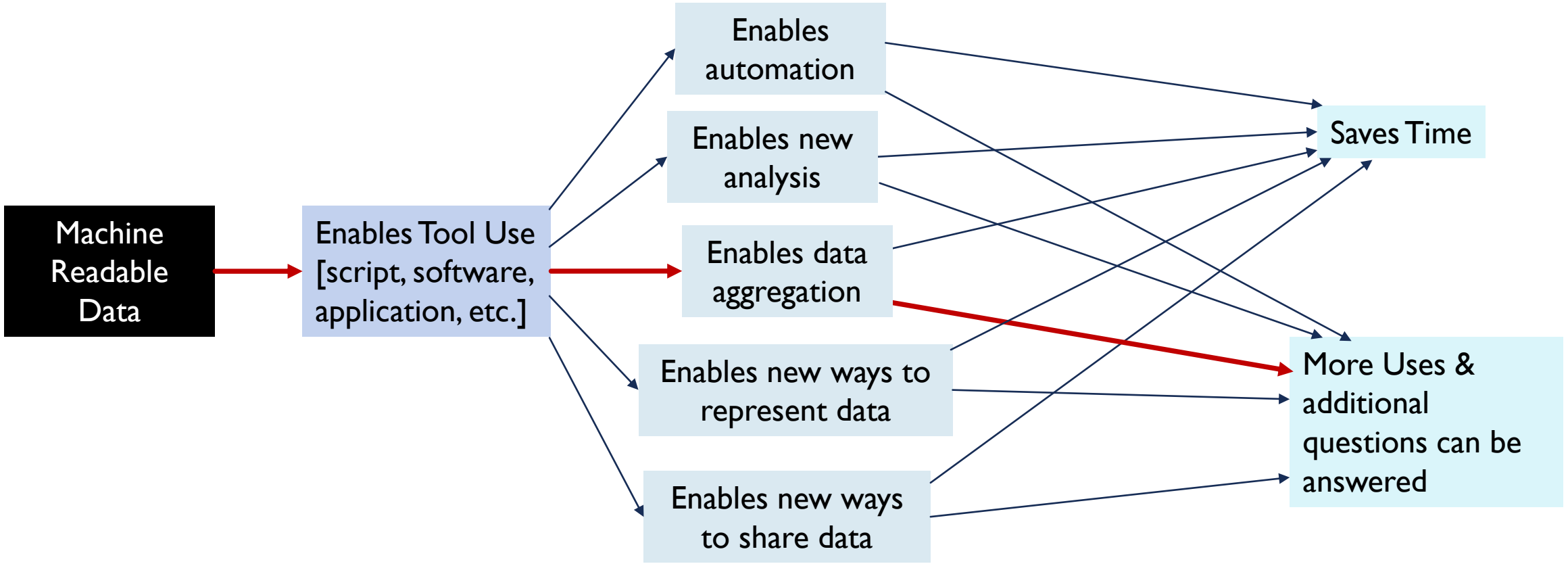
2) Extensive collaboration with procurement and quality engineering subject matter experts to make sure the technology, people, and processes can all work in sync to enable modern data analytics on hardware inspection data!

3) Final Deployment of working system & workflows

CONTENTS

These parts of the presentation are generic and not limited to hardware inspection data

WHY IS MACHINE READABILITY IMPORTANT?



Link to data.gov primer on machine-readable data

Red line represents the main driver for hardware inspection data modernization

DEFINITION:
WHAT DO WE MEAN BY MACHINE READABLE?

Defined in the 2019 Foundations for Evidence-Based Policymaking Act as….

- *"Data in a format that can be easily processed by a computer without human intervention while ensuring no semantic meaning is lost."*

DIFFERENT LEVELS OF 'DIGITALNESS'

Analog	Pseudo Digital	Machine Readable
Printed paper with tables	PDF with images of tables	CSV or JSON

- Should be able to read the actual content into other tools, ideally open-source ones, such that it is reusable in parts.

Idea from Matt Hall's blog post "is-your-data-digital-or-just-pseudodigital".

DEFINITION: "MACHINE READABLE" A SINGLE [CSV OR EXCEL] FILE

The central concepts of CSV dataset TIDYNESS:

- Each ROW is a separate observation or record
- Each COLUMN is a separate field
- Every CELL is only one piece of information

Whether it is JavaScript libraries, Tableau, Python packages, or Microstrategy….. Machines will expect the header row has the field labels and each row is a new observation.

Header row has field names

Each cell is a field value for that row

Columns = Fields

Rows = observations

Date	LastName	FirstName	Title	FileType
2021/04/07	Adrian	Andrew	Microlearning: data	pptx
2021/04/22	Gosses	Justin	How to store your data to enable modern data analytics: NASA's hardware inspection data	pptx

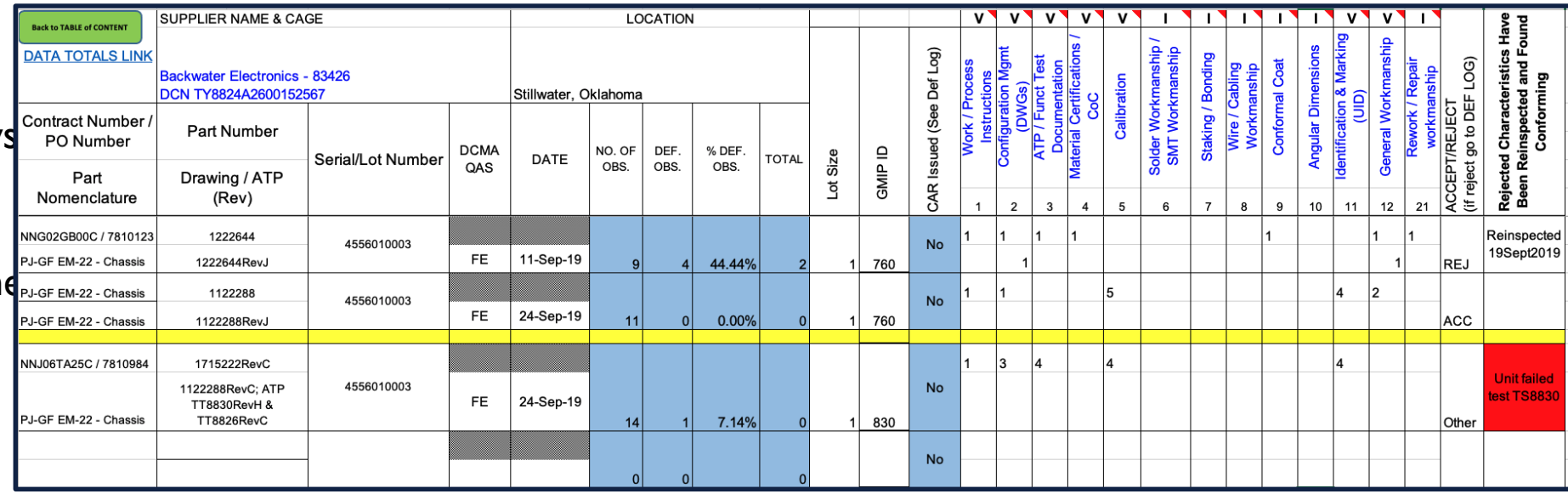
DEFINITION: ”MACHINE READABLE” DO’S & DON’TS

- Do not use white space on top, sides, or as dividing empty rows.
- Don’t embed meaning that you want to keep such that it only exists in formatting. For example, don’t color something red meaning “bad” but not have a cell that spells “bad” as well.
- No plots in the data sheet. Move them elsewhere.
- Don’t mix raw and calculated data in the same column. Calculated fields are better in separate tab or file.
- The column headers should contain only a unique name and [units], e.g. **Depth [m]**, **Porosity [v/v]**.
- No units in numeric data cells, only quantities. Record depth as **500**, not **500 m**. Put units in separate column or column name.
- Zero means zero, empty cell means no data.
- Avoid keys or abbreviations if possible.
- Try to use only one type of data per column: text OR numbers, discrete categories OR continuous scalars.

Sourced from Andrew Adrian & https://agilescientific.com/blog/2017/8/15/organizing-spreadsheets

BAD EXAMPLE TRANSFORMED INTO GOOD EXAMPLE

- Skipped Rows/Columns
- Merged Cells
- Single observations in 2 rows
- Multiple data in one cell
- Colors convey meaning
- Important data in sheet name
- Text/Numbers mixed



Back to TABLE of CONTENT / DATA TOTALS LINK	SUPPLIER NAME & CAGE			LOCATION								V	V	V	V	V	I	I	I	I	I	V	V	I			
	Backwater Electronics - 83426 DCN TY8824A2600152567			Stillwater, Oklahoma																							
Contract Number / PO Number Part Nomenclature	Part Number Drawing / ATP (Rev)	Serial/Lot Number	DCMA QAS	DATE	NO. OF OBS.	DEF. OBS.	% DEF. OBS.	TOTAL	Lot Size	GMIP ID	CAR Issued (See Def Log)	Work / Process Instructions	Configuration Mgmt (DWGs)	ATP / Funct Test Documentation	Material Certifications / CoC	Calibration	Solder Workmanship / SMT Workmanship	Staking / Bonding	Wire / Cabling Workmanship	Conformal Coat	Angular Dimensions	Identification & Marking (UID)	General Workmanship	Rework / Repair workmanship	ACCEPT/REJECT (if reject go to DEF LOG)	Rejected Characteristics Have Been Reinspected and Found Conforming
												1	2	3	4	5	6	7	8	9	10	11	12	21		
NNG02GB00C / 7810123	1222644	4556010003									No	1	1	1	1					1			1	1		Reinspected 19Sept2019
PJ-GF EM-22 - Chassis	1222644RevJ		FE	11-Sep-19	9	4	44.44%	2	1	760			1												REJ	
PJ-GF EM-22 - Chassis	1122288	4556010003									No	1	1			5						4	2			
PJ-GF EM-22 - Chassis	1122288RevJ		FE	24-Sep-19	11	0	0.00%	0	1	760															ACC	
NNJ06TA25C / 7810984	1715222RevC	4556010003									No	1	3	4		4						4				Unit failed test TS8830
PJ-GF EM-22 - Chassis	1122288RevC; ATP TT8830RevH & TT8826RevC		FE	24-Sep-19	14	1	7.14%	0	1	830															Other	
					0	0		0			No															

supplier_name	supplier_CAGE	location_city	location_state	part_number	serial_number	lot_size	GMIP_ID	drawing_number	drawing_number_revision
Frontier Electronic Systems	63812 DCN S4402A1609023	Stillwater	Oklahoma	PTPU-SM EM-2	2018010003	1	830	1722633	J

HOW TO ENSURE MACHINE READABILITY WHEN DEALING WITH MANY FILES AND NOT ALL IDENTICAL FIELDS?

"Be able to write one set of programmatic instructions that will always successfully aggregate the files into a single big file"

- Standardized set of fields
  - (*where possible*)
- Standardized definitions & definition capture
  - *(Of field definitions & which fields in which sheets*)
- Standardized placement of content
  - (Of data, definitions, and unexpected content!)
  - (Clear separation between data for analysis and human readable context with instructions)

WHAT WE'VE FOUND

Summary of the state of hardware inspection data and changes needed to enable data analytics

TENTATIVE CONCLUSIONS ON POTENTIAL SOLUTIONS

- *CAN NOT*: Mandate a single set of fields for all GMIP inspection forms, because:
  - Both the inspection form & the analytical capabilities need to support: [these explained more on next slide]
    - (1) pre-defined standard fields
    - (2) different pre-defined fields based on the type of inspection
    - (3) fields created by inspection owner
    - (4) place for non-expected information supplied by inspector or other parties such that it doesn't lower data quality.

- *CAN NOT ASSUME:* It is possible to translate pre-existing inspection data into a standardized set of fields:
  - Lack of field definitions means some translations will be guesses at best with high error rates.
  - Not all wanted fields will be recorded by inspector unless asked up front.

- *LIKELY CAN NOT ASSUME*: Everyone involved could be asked to log in and use a single application, because:
  - There should be the assumption that at least some information will still be sent at some point by files in email even if web applications are used as core part of eventual solution.

TYPES OF HARDWARE INSPECTION FIELD VARIANCE & HOW TO HANDLE AS TO ENABLE ANALYTICS

Groups of Fields	Who Creates	How to ensure aggregate analytics possible
▪ Uniform in all inspections	Mandatory & optional fields decided in advance by SME org	Maximize % of fields that are in these & relate to business questions! Additions okay, definition changes bad.
▪ Vary by Process & Standardized	Mandatory & optional fields decided in advance by SME org	Ensure these don't change much and are well defined terms that everyone understands. Additions okay, definition changes bad.
▪ Requested One-offs	Created by inspection requestor according to pre-defined data schema / methods	Make it easy for these to be well defined and recorded in places that are known in order to enable programmatic extraction. If they occur multiple times, move them to optional fields in light blue box.
▪ Unrequested Information	Populated by inspector in standardized location & way so as to not lower data quality	Make it easy for these to be well defined and recorded in places that are known in order to enable programmatic extraction.

PROTOTYPE COMPONENTS OF A SOLUTION

The next few slides describe:

The things we've built to explore the solutions space

These are not final products but more first pass artifacts. Future versions of them will likely be used in some way in a final solution.

SUMMARY OF OUR CURRENT APPROACH / PRODUCTS

As final product and user workflow is not clear at this point in time, we've focused on building reusable data products and working prototypes that allow exploring the solution space

1. **Field Schema:**
   - A standardized way to define hardware inspection fields according to several characteristics.
2. **First pass at standardized fields**:
   - This is based on analysis of historical data.
3. **Excel Template**:
   - A way to organize excel files (that could be adapted to a web application format) such that fields captured could be both standardized & variable, yet analytics still possible as field definitions and field placement are defined in the same place as the data.
4. **Data faker**:
   - A python package that leverages previous 3 items to fake large numbers of inspection reports. This will help SMEs see what is possible from aggregate data analysis and help test out field definitions, field schema, etc.
5. **Web application to create inspection forms**:
   - A working prototype of a web application that helps inspection requestors create inspection forms. Will use in conversations to understand current and possible user workflows.

SCHEMA FOR DEFINING FIELDS

Schema Fields that Describe Each Inspection Report Field (aka column)

- title
- singular_or_rows
- description
- examples
- type
- regex_pattern
- enum
- plain_language_validation
- dependency
- required_to_be_included_in_any_inspection_report
- required_to_be_filled_out
- who_fills_out
- array_sheetnames_with_this_field
- links_to_more_information

LEGEND

Yellow Background = Must be filled out for each field

Clear Background = optionally filled out

EXAMPLES OF STANDARDIZED MANDATORY INSPECTION FIELDS

title	singular_or_rows	grouping_of_fields	description	examples	type
Inspection_form_generation_number	singular	base	A unique number generated when the inspection form is created. I	202002111234.00	string
date_form_generated	singular	base	This will be generated for you and is the date the form was generat	"2020-11-27"	string
date_due_back_to_NASA	singular	base	The date the form is due completed back to NASA org that request	"2020-11-27"	string
NASA_program	singular	base	The name of the largest NASA program	"international Spac	string
NASA_name_of_largest_physical_entity	singular	base	The name of largest physical entity the hardware part will eventua	"International Spac	string
Requesting_NASA_org	singular	base	The name of the NASA organization requesting the hardware inspection. This won't be		string
Inspection_form_name	singular	base	A user provided name for the excel file that gets generated.	"test"	string
date_assigned_by_NASA_requestor	singular	base	The date the inspection request was sent to NASA procurement.	"2020-11-27"	string
date_sent_back_to_GMIP_once_complet	singular	base	The date GMIP received the completed inspection form.	"2020-11-27"	string
data_completed_and_sent_from_GMIP_t	singular	base	The date the inspection report is processed by GMIP and available	"2020-11-27"	string
GMIP_number	singular	base	unique identifier in GMIP system		string

EXCEL TEMPLATE STRUCTURE

Helps Ensure Machine Readable even if structure non constant

Data Captured From Here

Instructions	Definitions Sheets	Definitions Fields	InformationForAllSheets	Data_1	Data_2	CommentsAndState
Human-readable	Machine-readable	Machine-readable	Machine-readable	Machine-readable	Machine-readable	Human-readable
Contains instructions for how to fill out the inspection form.	Definitions for which fields are in which sheets	Definitions for each field in data sheets to the right	Data that applies across all data sheets in this file. Typically no rows but singular fields.	Tabular data, Each row is observation Each sheet is for either a day or an entity	Tabular data, Each row is observation Each sheet is for either a day or an entity	A place inspectors and others can put information that doesn't fit in other data sheets

Ensures data is machine-readable even if number and type of fields captured is variable

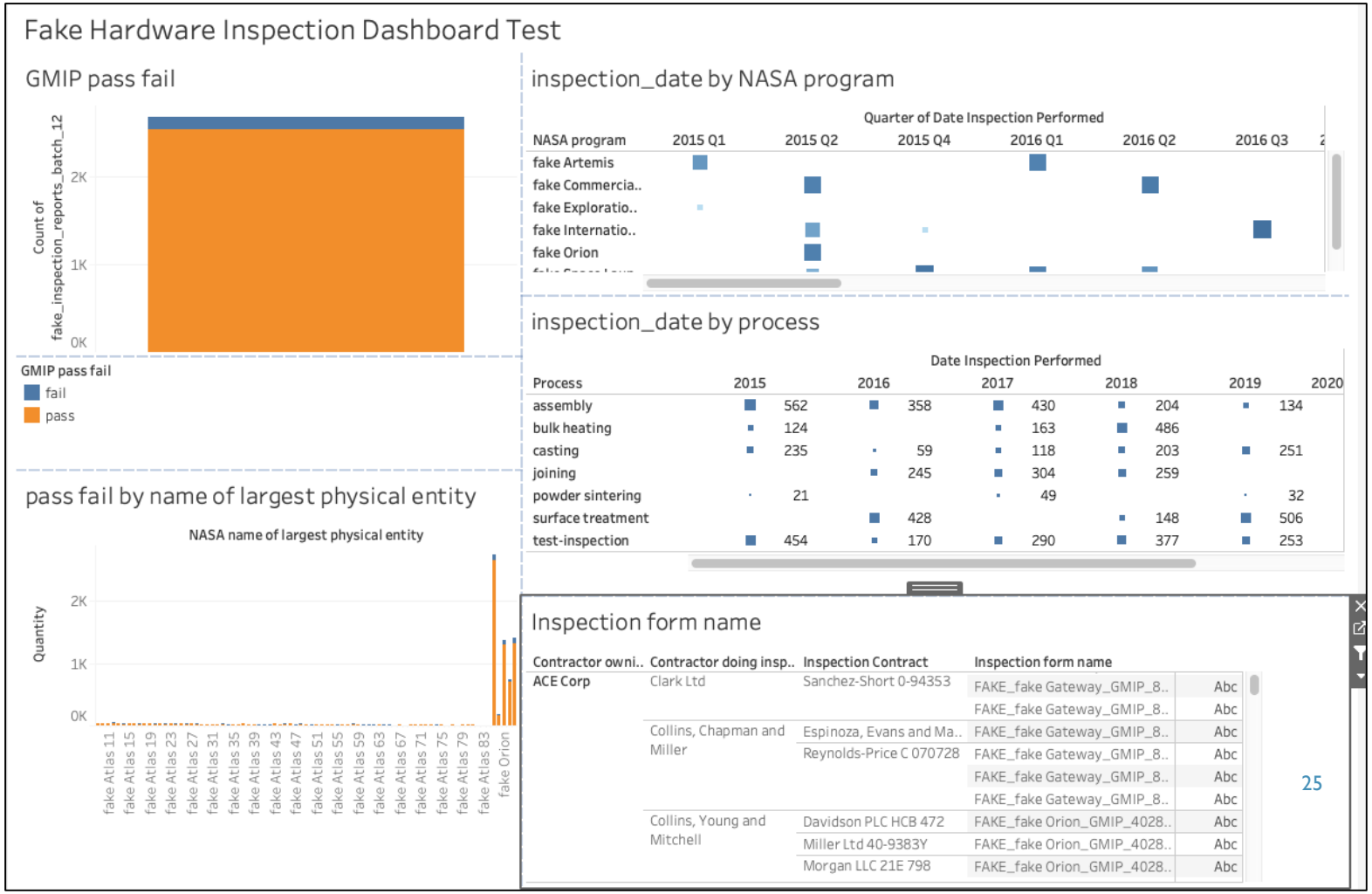
INSPECTION DATA FAKER:

1/2 MAKING SURE WE HAVE THE RIGHT DATA FIELDS BY EXPLORING WHAT QUESTIONS CAN BE ASKED

- Python package leveraging open source Faker package.
- Built on fields schema & excel template
- Generates X number of fake inspection data reports
- Enables prototyping of full analytics & visualization cycle

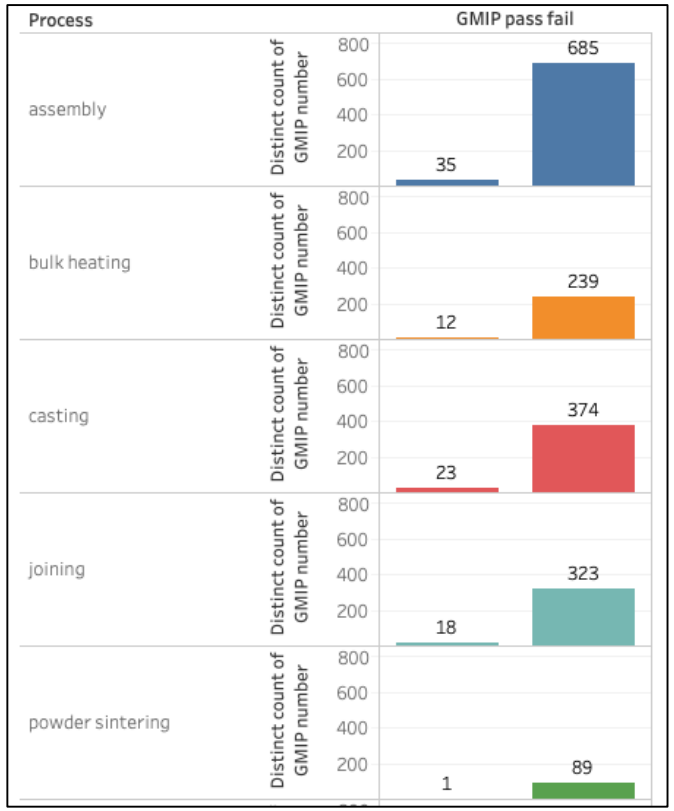
(currently using tableau for visualization as that was easy to throw together as example)

LINK TO REPO

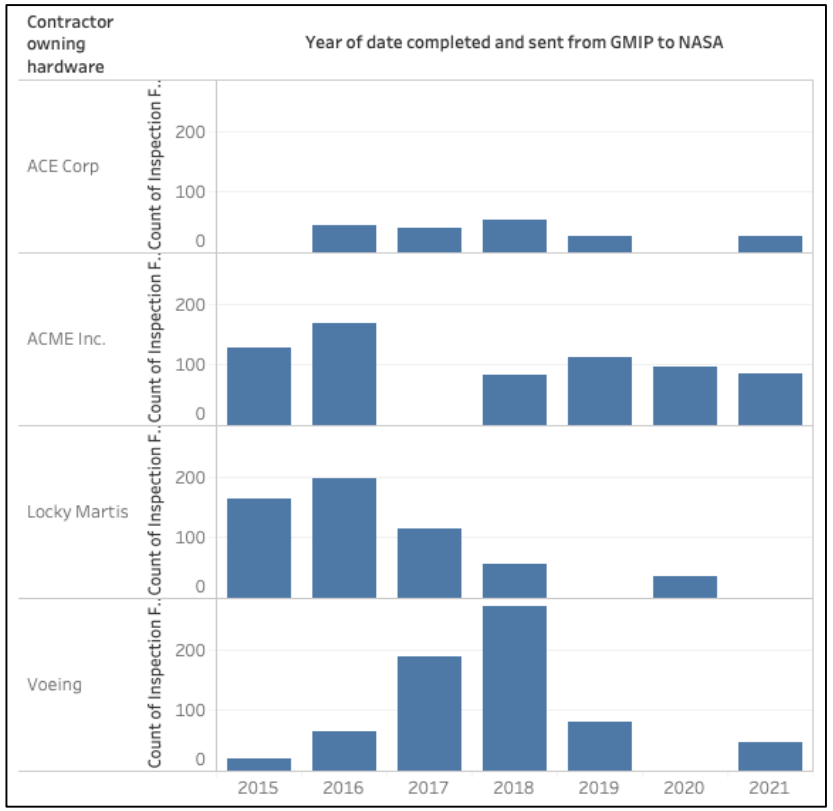


INSPECTION DATA FAKER:

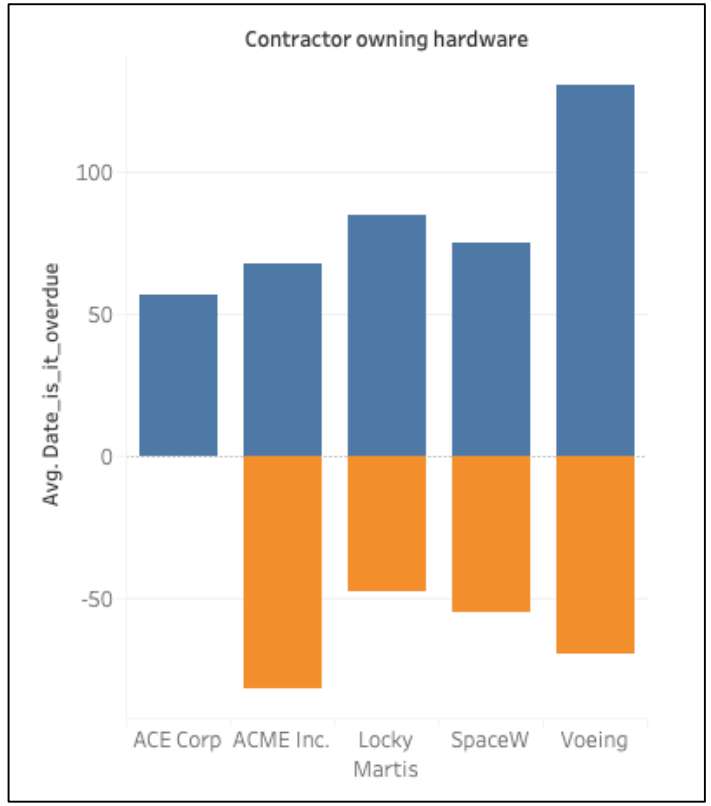
2/2 MAKING SURE WE HAVE THE RIGHT DATA FIELDS BY EXPLORING WHAT QUESTIONS CAN BE ASKED



How does individual GMIP failure rate vary by inspection process ____?



How has the count of inspection reports varied over time for contractor ____?



Is contractor ____ different than others in terms of how late or early reports come in ?

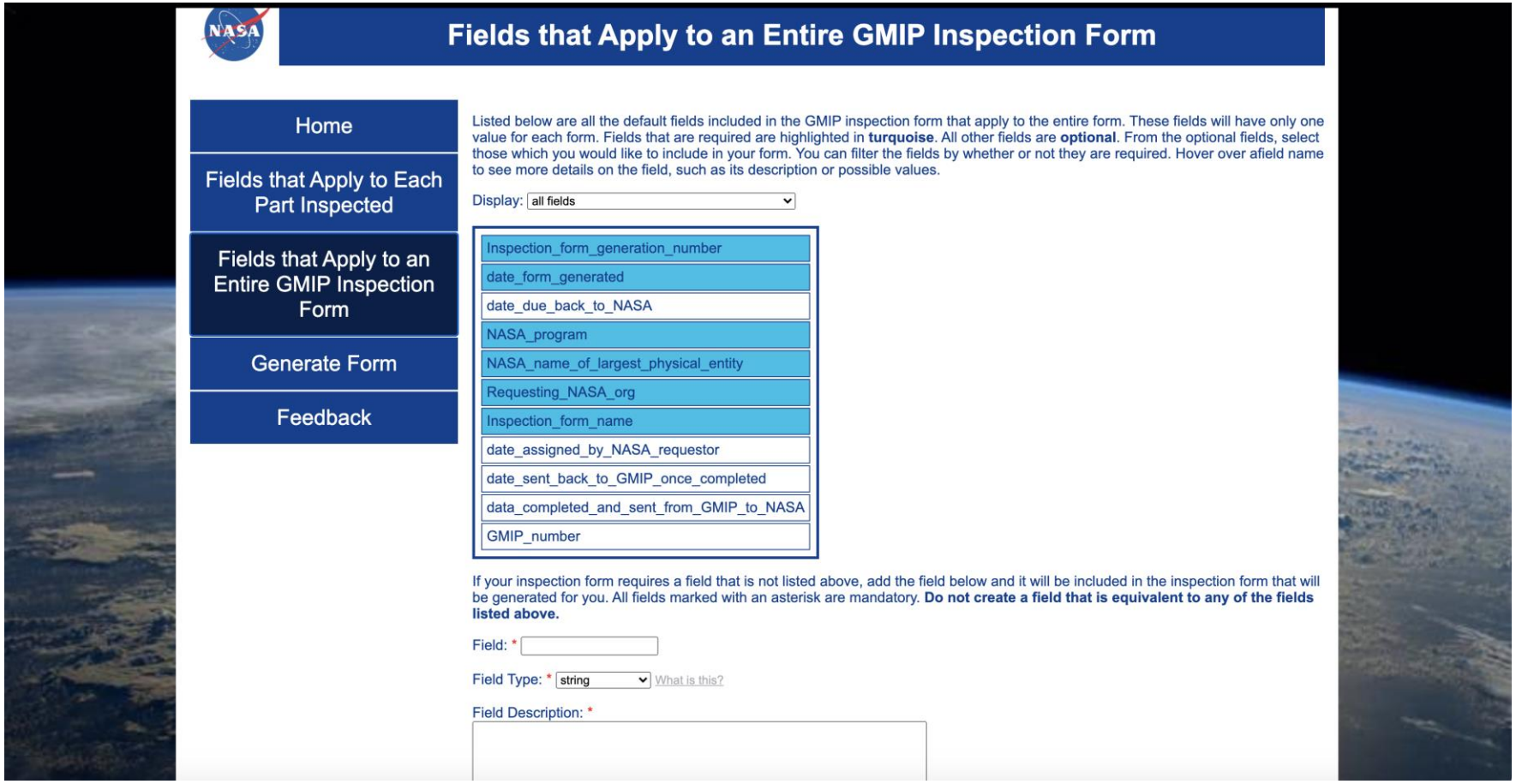
WEB APPLICATION TO CREATE INSPECTION FORMS:

As part of the solution, we propose inspection requesters be allowed to **generate their own inspection form using a website.**

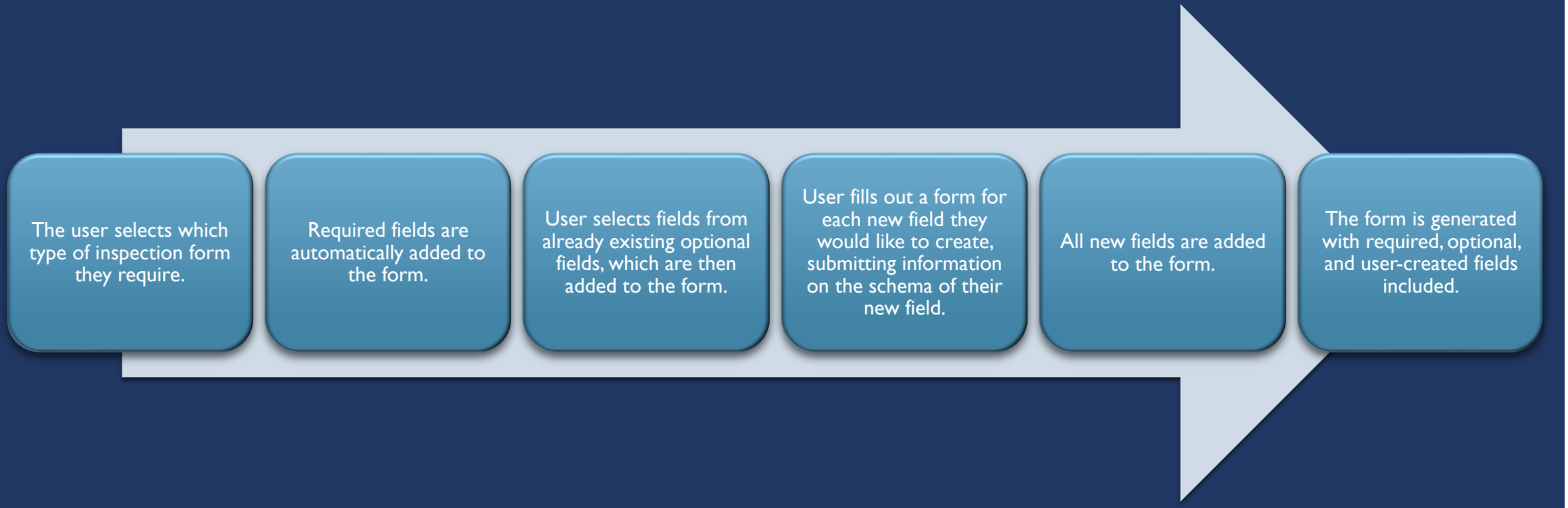
This approach ensures

1. Maximizes standardization of fields to extent practical
2. Let's inspection requestors specify what type of inspection (and resulting fields) apply to them.
3. Let's inspection requestors add in additional fields in a way that they are defined.
4. Creates an inspection form to be completed that is machine readable.

PROTOTYPE WEB APPLICATION TO CREATE INSPECTION FORMS: SCREENSHOT



USER WORKFLOW
FORM GENERATION PROCESS



MOVING FORWARD

What will be built?

What eventual system/product will you use?

MOVING FORWARD....

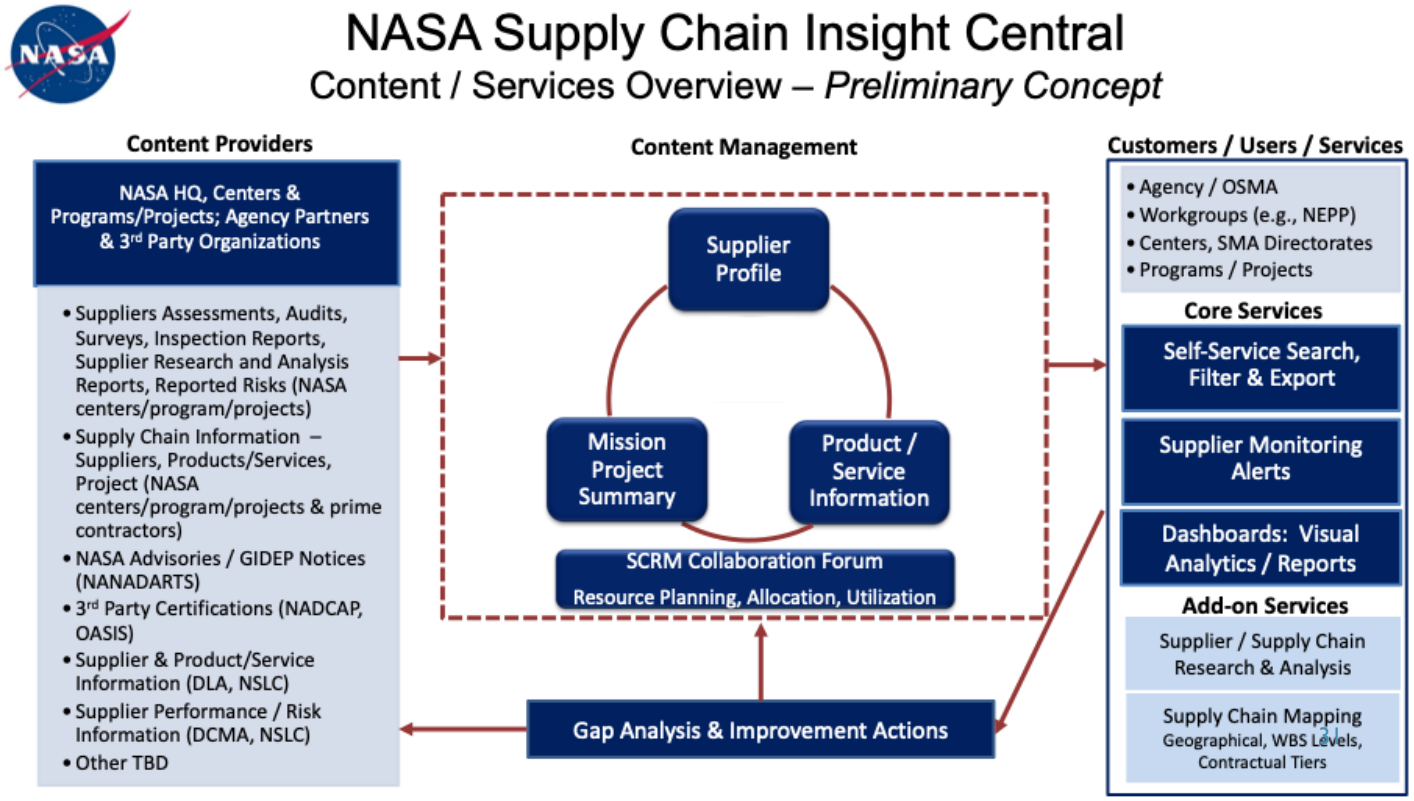
1. Field Standardization
2. User flow modeling
3. Build Final Applications
4. Deploy
5. Establish user documentation
6. Continuously evolve

Elizabeth and Justin's work on this project will shrink

Bulk of work will now be done by Goddard META team & procurement.

Actual Deployed Applications will be tied to SCIC

Diagram of upcoming SCIC



QUESTIONS?

EXAMPLE: HOW FIELDS VARY ACROSS INSPECTION REPORTS

	Uniform Across Inspections		Process Specific		Variances Important to Capture in Known Manner	
Inspection Type	Always Mandatory	Always Optional	Mandatories for specific Process	Optional for specific process	Requested One offs	Unrequested Information
Assembly, *Company A*						
Assembly, *company B*						
Assembly & Finish, *company B*						
Assembly & Finish, Company C						
Finish, *company A*						
Finish, *company B*						
Finish, *company C*						

LEGEND

Each dot is a field

Same colored dot in a column represents the same field.

 in "always mandatory" might represent date an inspection request sent out &  in "mandatories for specific process" might represent are all mandated parts presents.