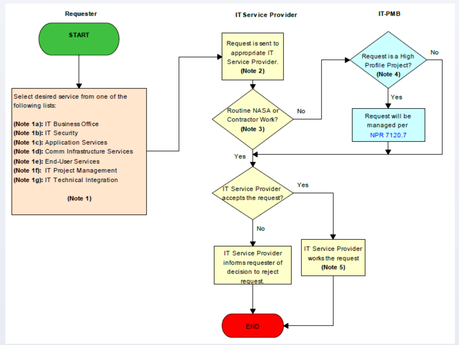
Migrating a Website to Sitecore

John McCardle

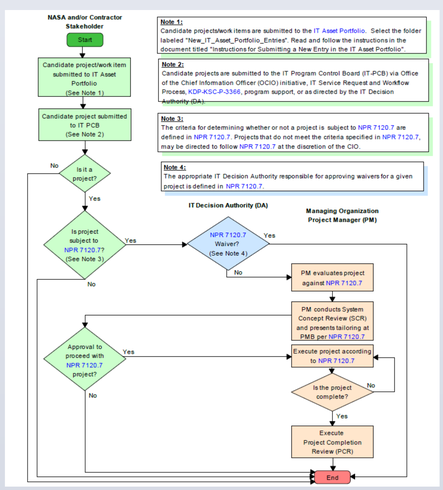
IT-C2 – Application Engineering and Operations Branch

IT Processes



From KDP-KSC-P-3366, IT Service Request and Workflow Process

An IT Service Request is the starting point for any need, big or small. After the request is analyzed, it will be routed to the relevant IT Branch or Contractor.



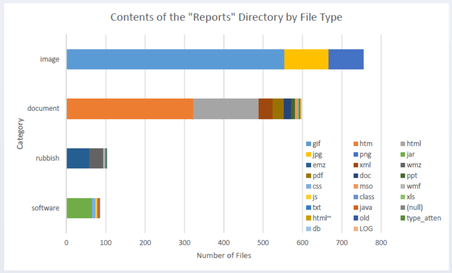
From KDP-P-3355, Project Management

An IT request of a certain complexity will not be assigned immediately, but instead will become a project. Projects are treated as unique events which must be planned, approved, executed, and reviewed according to procedure. Projects allow IT to address needs for which there is no established procedure.

In IT-C2, our projects are conducted in sprints: a 2 or 3 week period of development which always ends in a prototype or product demonstration.

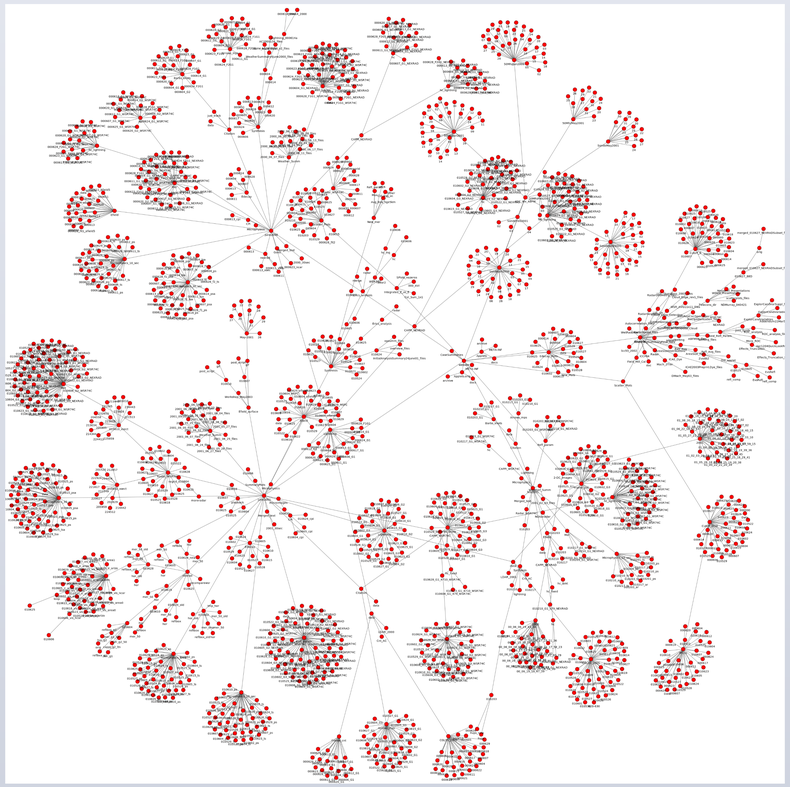
Website Migration Analysis

Nothing is permanent. Though digital products may have an immaterial quality to them, software and websites have a shelf life. Websites only exist to support missions, and when a mission outlasts its web technologies, a *migration* is necessary. In this process, we assess the website and move it to a new platform. For my first assignment as a NASA civil servant, I am assessing a website for migration to Sitecore.



Reports Detected by Web Scraper

My web scraper program detected over 1,500 files that went undetected by earlier site analysis.



Web Scraping is the technique of using a computer program other than a web browser to access websites, in order to collect all or some of the data. In this case, I used a program called a web scraper or *spider* to collect a complete listing of the files and directories of one KSC website. This information helps our customer to make determinations about what files are still in use, which files need to be archived, and which can be safely discarded.

```python
def analyse_link(href):
"""Determines what file type or link a URL points to."""
        if href.endswith('/'):
                ltype = "dir"
        elif href.startswith('mailto:'):
                ltype = "email"
        else:
                ltype = "file"
        return ltype

def links_on_page(url):
"""Returns a list of bare URLs on the given page. Full URL required."""
        r = requests.get(url)
        soup = BeautifulSoup(r.text, 'html.parser')
        #print(f"page {url} {len(r.text)} bytes")
        return [l['href'] for l in soup.find_all('a') if
        analyse_link(l['href']) == 'dir']

def isolated_page(url):
"""Splits a URL into components and returns the final portion and the
directory depth. Either full or relative URLs will work, but don't compare
the depth of full and partial"""
        isolate = [s for s in url.split('/') if not len(s) == 0]
        if len(isolate) == 0: return ('/', 0)
        else: return (isolate[-1], len(isolate))

def analyse_page(start_url, base_page):
"""Recursively follows relative URLs on a page to map out all
subdirectories."""
        url = base_page + start_url
        this_page, depth = isolated_page(start_url)
        directory = {}
        links = links_on_page(url)
        #print(f"page: {url} -> '{this_page}' links: {len(links)}")
        for link in links:
                l, d = isolated_page(link)
                if d < depth:
                        #print(f"Skipping link to {l}")
                        continue
                #print(f"found sublink[{d}] to {l}")
                directory[l] = analyse_page(link)
        return directory

import json
map_json = json.dumps(page_map, indent=4) #indent for pretty printing
with open("page_map.json", "w") as f:
        f.write(map_json)
```

Excerpt of Python Code

This web scraper, written in Python, does not take much code. The secret to its effectiveness is running the code recursively over the target website's 1,800+ subdirectories.

Website Directory Tree

This tree of a KSC website was generated with data gathered by the web scraper. Every vertex represents a directory, and the connecting lines indicate a directory's contents. The web scraper recursively explores every subdirectory until the entire site has been indexed. This particular website had 1,699 subdirectories.

Following my investigation, we were able to categorize the contents of the website to separate relevant content from historical and out of date content. For a site of this magnitude, it's crucial that we only spend effort migrating content that is still of value to the organization.

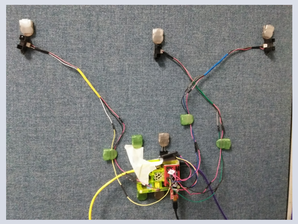
Tools

In Information Technology, we occasionally use screwdrivers and mallets, but more frequently our work calls for virtual tools. It takes software to make software.

- Python: A quick and easy scripting language that is often used to prototype new ideas.
- Sitecore: A Content Management System (CMS) that hosts many of KSC's websites using standard templates.
- C#: A programming language developed by Microsoft. Programs written in C# have high performance for performing intense computations, and can transfer information to a database or your web browser.
- Angular: A Javascript framework for making web applications. This is the tool of choice when a Content Management System doesn't have the flexibility to display dynamic data or process user input.
- Excel: The common spreadsheet is a critical tool in visualization and decision making.

Fun

The office can be a quiet environment: so quiet that any minor disturbance becomes an alarming experience. To prevent my co-workers from startling me when entering my cubicle, I built this entry detector. Different sound effects are played for entry and exit.



Three infrared distance sensors scan for objects passing between them and the opposite wall. When the beam is broken, software tracks the sequence of interruptions to determine the direction of travel.

Acknowledgements

I'd like to thank my mentor, Laurie Brown, for guiding me through my first weeks at NASA and showing me the ropes of Sitecore. Additionally, everyone in IT-C2 has earned my unending gratitude for their patience with my numerous questions.