Attachment #3: Blue Angel Rules

Two tests were applied by Blue Angel to determine if a document was in scope (i.e. considered to be an EPA publication). The first was to determine if a document was considered an in-scope publication, the second was to determine if the document is an EPA publication.

Rules to Determine if a Document is a Publication

These rules apply a test to see if a document is an in-scope publication. A document is considered to not be an in-scope publication if it meets any of the following criteria.

- **NP_UnsupportedType:** The document type is not supported.

- **NP_Abstract**: Indicates that a document is an abstract. This test checks if any of the following conditions are met:
  - The Subject or Title metadata field contains the word “abstract”
  - The Front Matter metadata field contains at least five (5) of the following words and phrases: “Abstract:”, “Citation:”, “Contact:”, “Division:”, “Branch:”, “Product Type:”, “Presented:”, “Related Entries:”

- **NP_Agenda:** Indicates that a document is a conference agenda. This test checks if all of the following conditions are met:
  - The Front Matter metadata field contains at least five (5) of the following words and phrases: “agenda”, “break”, “call to order”, “conference”, “cost”, “goal”, “goals”, “hotel”, “lodging”, “lunch”, “luncheon”, “master of ceremonies”, “meal”, “meals”, “opening comments”, “papers”, “presentation”, “presentations”, “registration”, “seminar”, “seminars”, “session”, “sessions”, “speaker”, “speakers”, “topic”, “track”, “travel information”, “welcome”, “workshop”
  - The First 250 Words, Subject, and Title metadata fields do not contain the word “proceedings”

- **NP_ConsentForm:** Indicates that a document is a consent form. This test checks if the Front Matter metadata field begins with the phrase "consent for"

- **NP_Docket:** Indicates that a document is a docket publication. This test checks if theDescription, Subject, or Title metadata field contains the phrase “Docket No.”

- **NP_Draft:** Indicates that a document is a draft. This test checks if any of the following conditions are met:
  - The Subject or Title metadata field begins or ends with the word “draft”
  - The First 250 Words metadata field contains the word “draft”

- **NP_Form:** Indicates that a document is a form. This test checks if the First 250 Words metadata field contains at least one of the following phrases: “amendment of solicitation”, “for instructions”, “For Sample Use Only”, “modification of contract”, “see instructions”, “type or print all information”

- **NP_Fragment:** Indicates that a document is a document fragment. This test checks if the Description or Subject metadata field contains at least one of the following phrases: “extracted page”, “extracted pages”, “from the”

- **NP_Instructions:** Indicates that a document is a set of form instructions. This test checks if the Front Matter metadata field begins with the phrase "instructions for"

- **NP_InternalSummaryMemo:** Indicates that a document is an internal summary memorandum. This test checks if the First 250 Words metadata field contains at least two of the following words and phrases: “Action:”, “Agency:”, “RFIP No.:”, “Summary:”, “Title:”

- **NP_Letter:** Indicates that a document is a letter. This test checks if any of the following conditions are met:
    - The First 250 Words metadata field contains the phrase "Dear <1-4 words>:" or “Dear <1-4 words>,” where <1-4 words> can be any set of one to four words.
    - The First 250 Words metadata field contains the phrase “letter from”

- **NP_MeetingAnnounce:** Indicates that a document is a meeting announcement. This test checks if all of the following conditions are met:
    - The Subject or Title metadata field contains the phrase “public meeting”
    - The First 250 Words, Subject, and Title metadata fields do not contain the word “proceedings”

- **NP_Memo:** Indicates that a document is a memorandum. This test checks if all of the following conditions are met:
    - The First 250 Words metadata field contains any of the following words: “memo”, “memorandum”
    - The First 250 Words metadata field contains at least two of the following words and phrases: “Attendees:”, “Date:”, “Date and Time:”, “From:”, “Location:”, “Re:”, “Subj:”, “Subject:”, “Time:”, “To:”
    - The First 250 Words, Subject, and Title metadata fields do not contain any of the following phrases: “memorandum of understanding”, “memorandum of agreement”, “memorandum of intent”

- **NP_MemoOfUnderstanding:** Indicates that a document is a memorandum of understanding. This test checks if the Subject or Title metadata field contains at least one of the following phrases: “memo of understanding”, “memorandum of agreement”, “memorandum of intent”, “memorandum of understanding”, “MOU”

- **NP_MetadataRecord:** Indicates that a document is a metadata record. This test checks if the Front Matter metadata field begins with the phrase "metadata record"

- **NP_Minutes:** Indicates that a document is a meeting minutes. This test checks if all of the following conditions are met:
  - The First 250 Words metadata field contains at least one of the following phrases: “Conference Call Summary”, “meeting minutes”, “meeting summary”, “public meeting”, “Stakeholders Meeting”, “summary meeting”, "summary minutes”
  - The First 250 Words, Subject, and Title metadata fields do not contain the word “proceedings”

- **NP_PurchaseOrder:** Indicates that a document is a purchase order. This test checks if the First 250 Words metadata field contains the phrase “purchase order”

- **NP_Readme:** Indicates that a document is a readme file. This test checks if the Front Matter metadata field begins with the word "readme"

- **NP_Solicitation:** Indicates that a document is a solicitation. This test checks if the First 250 Words metadata field contains at least one of the following phrases: “solicitation, offer, and award”, “this contract is a”, "type of solicitation”

- **NP_SOW:** Indicates that a document is a statement of work. This test checks if the First 250 Words metadata field contains the phrase “Statement of Work”

- **NP_SurveyForm:** Indicates that a document is a survey form. This test checks if any of the following conditions are met:
  - The Keywords metadata field contains at least one of the following words: "form", "questionnaire", "survey"
  - The First 250 Words metadata field contains the word “questionnaire”

- **NP_Testimony:** Indicates that a document is a testimony. Testimony is not considered a publication, as it would be included with the hearing and as such would be duplicative. This test checks if the Front Matter metadata field begins with any of the following words and phrases: “statement of”, “testimony”

Otherwise, a document is considered to be a Publication if it meets any of the following criteria:

- **QP_EnglLink1:** The Source Link Language is English and the Source Link Words imply that the document is a publication

- **QP_First250Words:** The First 250 Words imply that the document is a publication

- **QP_Funded:** The Front Matter contains the text along the lines of “publication was funded by grant or contract”

- **QP_ISBN:** An ISBN is found in the Front Matter or End Matter
- **QP_ISSN:** An ISSN is found in the Front Matter or End Matter
- **QP_SpanLink1:** The Source Link Language is Spanish and the Source Link Words imply that the document is a publication

Otherwise, the document is considered to not be a Publication.

Rules to Determine if a Publication is an EPA Publication

A document is deemed to be an EPA publication if all of the following criteria are not met:

- **QE_CFR:** The First 250 Words metadata field contains the string "CFR". Note that this is a string match and not a word or phrase match.
- **QE_CongressionalRecord:** The Subject or Title metadata field contains the phrase "congressional record"
- **QE_EPACDER:** The Front Matter or End Matter metadata field contains text referencing the EPA's Central Data Exchange Registration
- **QE_EPADirectory:** The Front Matter metadata field contains full text referencing the Environmental Protection Agency at least eight (8) times.
- **QE_FedReg:** The document is associated with the Federal Register
- **QE_NonUSEPA:** Indicates that a document is a non-U.S. E.P.A. publication. This test checks if any of the following conditions are met:
  - The Subject or Title metadata field contains any of the Agency or Abbreviation values found in Appendix: State Environmental Agencies
  - The Front Matter metadata field begins with any of the Agency or Abbreviation values found in Appendix: State Environmental Agencies
- **QE_TitleCFR:** The Title metadata field contains the word "CFR", the word "C.F.R.", or the phrase "Code of Federal Regulations"
- **QE_TitlePublicLaw:** The Title metadata field contains the phrase "Public Law"

AND any of the following criteria ***are*** met:

- **QE_AuthorEPA:** The Author metadata field contains full text referencing the Environmental Protection Agency

- **QE_DescriptionEPA:** The Description metadata field contains full text referencing the Environmental Protection Agency

- **QE_EndEPA:** The End Matter metadata field contains full text referencing the Environmental Protection Agency (see Algorithm EPATextFull).

- **QE_FrontEPA:** The Front Matter metadata field contains full text referencing the Environmental Protection Agency

- **QE_NonSeedEPA:** All of the following criteria are met:
    - The publication is not from a Seed URL
    - The Author, Description, Subject, or Title metadata field contains full text
    - referencing the Environmental Protection Agency, or abbreviated text referencing the EPA
    - The End Matter or Front Matter metadata field contains full text referencing the Environmental Protection Agency

- **QE_SubjectEPA:** The Subject metadata field contains full text referencing the Environmental Protection Agency

- **QE_TitleEPA:** The Title metadata field contains full text referencing the Environmental Protection Agency