Template Version Number: 01-2021

U.S. Department of Commerce
U.S. Patent and Trademark Office

Privacy Threshold Analysis
for the
Open Data-Big Data Master System (OD-BD MS)

U.S. Department of Commerce Privacy Threshold Analysis

USPTO Open Data-Big Data Master System (OD-BD MS)

Unique Project Identifier: PTOC-034-00

Introduction: This Privacy Threshold Analysis (PTA) is a questionnaire to assist with determining if a Privacy Impact Assessment (PIA) is necessary for this IT system. This PTA is primarily based from the Office of Management and Budget (OMB) privacy guidance and the Department of Commerce (DOC) IT security/privacy policy. If questions arise or further guidance is needed in order to complete this PTA, please contact your Bureau Chief Privacy Officer (BCPO).

Description of the information system: *Provide a brief description of the information system.*

The E-Government Act of 2002 defines "information system" by reference to the definition section of Title 44 of the United States Code. The following is a summary of the definition: "Information system" means a discrete set of information resources organized for the collection, processing, maintenance, use, sharing, dissemination, or disposition of information. See: 44. U.S.C. § 3502(8).

The Open Data/Big Data (OD-BD) master system consists of subsystems which support the Big Data Portfolio. OD-BD MS resides on the UACS platform, which employs IaaS and PaaS services from AWS and is located at USPTO Headquarters located at 600 Dulany Street Alexandria, Virginia 22314 ("IT EAST" Environment).

BDR:

The Big Data Reservoir provides USPTO employees a Big Data platform in which they can view records and associated metadata in one location. The Big Data Reservoir (BDR) is a Hadoop Distributed File System (HDFS) infrastructure used to perform advanced analytics on disparate data sets consisting of structured and unstructured data in order to gain insights and develop models. System users are USPTO Internal Users. BDR is a large repository for structured and unstructured data. Models and algorithms are developed with the BDR data to provide insights to PTO executives. Dashboards, search functionality, and visualizations provide users the ability to view the BDR data.

BDR-TQR:

In addition to the BDR Portal, the BDR also provides the TQR Portal. The TQR Portal provides quality reviewers with a centralized location to view the Dockets that are in the queue for review and additional features that include reviewing Trademark Review forms and completing necessary actions, final and non-final. System users are USPTO Internal Users.

BDR-CPC:

CPC is used to automatically classify patent documents. Users can place input .csv file by using SFTP, which contains number of application IDs. By using this input file, BDR AI API gets the contractor data from the CPC OracleDB and machine data from the BDR AI for corresponding application IDs and stores the data in two csv files. Users can compare contractor data and machine data, by giving application ID in the WEB. System users are USPTO Internal Users.

BDR-PTAB:
PTAB uses the BDR framework to gather data from PTAB E2E Oracle DB (PALMGP) and also from two OPSG REST APIs. Newly populated data in Oracle DB is collected by using Delta processing and stored in BDR HIVE/HDFS locations. The entire hive table's data is stored in SOLR index (Public), Elastic Index, and users can easily search the data based on a particular attribute. System users are USPTO Internal Users.

Address the following elements:

a) *Whether it is a general support system, major application, or other type of system*
OD BD MS is a major application.

b) *System location*

USPTO Headquarters located at 600 Dulany Street Alexandria, Virginia 22314 ("IT EAST" Environment).

c) *Whether it is a standalone system or interconnects with other systems (identifying and describing any other systems to which it interconnects)*

Information Dissemination Support System (IDSS): supports the Trademark and Electronic Government Business Division, the Corporate Systems Division (CSD), the Patent Search System Division, the Office of Electronic Information Products, and the Office of Public Information Services.

Trademark Processing System – External System (TPS ES): provides customer support for processing Trademark applications for USPTO.

Patent End to End (PE2E): provides examination tools for Central examination unit to track and manage the cases in this group and view documents in text format.

Trademark Next Generation (TMNG) - provides support for the automated processing of trademark applications for the USPTO.

Trademark Processing System – Internal System (TPS-IS): provides support for the automated processing of trademark applications for the USPTO. TPS-IS includes eleven applications that are used to support USPTO staff through the trademark review process.

Patent Capture and Application Processing System – Examination Support (PCAPS-ES): processes, transmits and stores data and images to support the data-capture and conversion requirements of the USPTO to support the USPTO patent application process.

Patent Capture and Application Processing System – Capture and Initial Processing (PCAPS-IP): captures patent applications and related metadata in electronic form; processing applications electronically; reporting patent application processing and prosecution status; and retrieving and displaying patent applications. PCAPS-IP is comprised of multiple Automated Information Systems (components) that perform specific functions, including submissions, categorization, metadata capture, and patent examiner assignment of patent applications.

Enterprise Software Services (ESS): provides an architecture capable supporting current software services as well as provide the necessary architecture to support the growth anticipated over the next five years.

d) The purpose that the system is designed to serve

The system is designed to serve as the enterprise platform for advanced analytics

e) *The way the system operates to achieve the purpose*

The Open Data/Big Data (OD/BD) master system consists of subsystems which support the Big Data Portfolio. OD/BD resides on the UACS platform, which employs IaaS and PaaS services from AWS. The current subsystem under this master system consists of Big Data Reservoir (BDR), Big Data Reservoir TQR (BDR-TQR), BDR Cooperative Patent Classification (BDR CPC), BDR Patent Trial and Appeals Board (BDR-PTAB), Developer Hub (DH) and Developer Hub Assignment Search (DH-AS). The system is designed to serve as the enterprise platform for advanced analytics.

f) A general description of the type of information collected, maintained, used, or disseminated by the system

a. Patent Data
- *i.* Publicly disseminated data (PG Pubs/Grants, PTAB decisions)
- *ii.* Patent Attributes (PALM)
- *iii.* Patent Office Actions (P-ELP), PATI-CDC data

b. Trademark –
- *i.* FAST – Tagged paragraphs
- *ii.* TRM – Trademark application information

iii. CMS – Trademark mark graphic
iv. BDR-TQR – TQR UI captures Trademark review information

g) *Identify individuals who have access to information on the system*

Users who have access to the system include Patent Executives, Trademark Reviewers, Data scientists, and System Administrators/Operators.

h) *How information in the system is retrieved by the user*

BDR is a large repository for structured and unstructured data. There is the compute tier, where the data is loaded, compared for public versus private status, and analyzed according to data science principles. There is the analysis tier, where data scientists combine the real-world problem-solving techniques from Patent Examiners with the formulae and hypothesis of the Data Science field. The Visualization tier that provides the users with a place to view the analysis and the underlying data that helps to create it. Finally, in the storage tier, the system retains raw, merged and transformed data, disseminates between public and private Patent applications and segregates them. Dashboards, search functionality, and visualizations provide users the ability to view the BDR data.
Developer Hub (DH) uses an N-tier architectural design pattern that separates the processing logic into distinct processing layers. The system is logically divided into six major subsystems:

- Access Layer: The access layer includes client web browsers and applications. Browser-based users can access Developer Hub web front and its contents. Users can also view DS- API Swagger pages and perform searches on data for various data sets.
- Web Server Layer: This layer hosts Apache Web servers. To follow USPTO EA standards, Apache is configured as the web server in front of the Wildfly server. The Web Server layer serves two purposes—presenting Developer Hub's static and dynamic content, and receiving and responding to DS- API web services calls (HTTP Get/Post messages). The Application Load Balancer routes the web services calls to Wildfly, which hosts the DS- API Web services. Apache Web Server is the server and uses AWS Elastic Load Balancing (ELB) for load balancing applications.
- Application Server Layer: This layer uses Wildfly application server to host various Springboot Web Services such as user authentication, email subscription / notification, ETL process and data synchronization. The Wildfly servers are configured in cluster; if a server goes down, subsequent user requests can be forwarded to a different server.

- Search Layer: In BDR, we are utilizing ElasticSearch to provide search capabilities against various data-sets. In DH, we are using SolrCloud to provide the same search capabilities.
- Data Layer: The Data Layer is responsible for providing access to the data from various sources, such as Drupal Relational Database (RDS), DS- API Relational Database (RDS), Unstructured Events Data (AWS S3).
- Infrastructure Layer: This layer provides user registration, authentication and authorization using Okta.

The Developer Hub Assignment Search (DH-AS) system indexes patent assignment records and allows them to be searchable by the public. To accomplish this, the system writes the internal records as files and transfers them to a receiving file system. A process monitors this file system and sends the records to the search system for indexing. Once complete with indexing, the whole file is transferred to another file system. If any errors occur, a third file system receives the file.

i) How information is transmitted to and from the system

BDR: Information is transmitted through batches, service calls, and user entry (BDR-TQR feature). All transmissions and retrieval of information are performed within the USPTO network and do not exceed the internal network boundary.

The BDR application employs a multilayered design approach. This approach gives modularity to the system. The following sections explain in high level, how each layer is comprised. The design principle of the BDR aims to have a tiered approach to the application. This way every component of the ecosystem is more easily understood and viewed independently. In this platform, there is ingestion, where the data is ingested from existing software resources. There is the compute tier, where the data is loaded, compared for public versus private status, and analyzed according to data science principles. There is the analysis tier, where data scientists combine the real-world problem-solving techniques from Patent Examiners with the formulae and hypothesis of the Data Science field. The Visualization tier that provides the users with a place to view the analysis and the underlying data that helps to create it. Finally, in the storage tier, the system retains raw, merged and transformed data, distinguishes between public and private Patent applications, and segregates them.

Developer Hub (DH):
The DH system provides USPTO public data (such as patents, trademarks, and events data) via a set of Web Services APIs for the consumption of the developer community. These APIs will be

developed and maintained by various divisions within USPTO and will be accessible through a USPTO web UI named Developer Hub, or Davent Hub (DH) System Name.

The system provides access to USPTO public content through the use of APIs (application programming interface). It has been determined that DH does not process PII/BII information, and it is categorized as a low risk system. The DH web application is deployed on the Amazon Web Services (AWS) Cloud platform. Users include: General Public, System Development Staff, Tableau Public Users, EC2 Server Accounts, Drupal Admin User via RBAC, and System Administrators.

Developer Hub Assignment Search (DH-AS):
DH-AS is responsible for indexing patent and trademark assignment records, which allows them to be searched by the public. To accomplish this, the internal records are written as files and transferred from AHD to a receiving file system. DH-AS is hosted on an AWS Public Cloud using the IaaS Service Model. It has been determined that DH-AS does not process PII/BII information, and it is categorized as a low risk system. The DH web application is deployed on the Amazon Web Services (AWS) Cloud platform. Users include: PTONet internal users - Assignment Historical Database (AHD), Assignment Services Branch, USPTO personnel such as patent examiners and support staff, Public Search Facilities staff members, and SOLR administrators.

AS provides public access via Amazon's Web Service Cloud the capability for external users of the USPTO as well as public users in the USPTO public search rooms (with access to the Internet) to query issued patent or published application patent assignment data and/or pending or registered trademark assignment data. The AS web application is deployed to the middleware environment running under Apache web servers and is available to external customers/users of the USPTO (outside of PTONet) via the Internet.

Questionnaire:

1. Status of the Information System

1a. What is the status of this information system?

☐ This is a new information system. *Continue to answer questions and complete certification.*

☐ This is an existing information system with changes that create new privacy risks. *Complete chart below, continue to answer questions, and complete certification.*

Changes That Create New Privacy Risks (CTCNPR)					
a. Conversions	☐	d. Significant Merging	☐	g. New Interagency Uses	☐
b. Anonymous to Non-	☐	e. New Public Access	☐	h. Internal Flow or	☐

Anonymous				Collection	
c. Significant System Management Changes	☐	f. Commercial Sources	☐	i. Alteration in Character of Data	☐
j. Other changes that create new privacy risks (specify):					

☐ This is an existing information system in which changes do not create new privacy risks, and there is not a SAOP approved Privacy Impact Assessment. *Continue to answer questions and complete certification.*

☒ This is an existing information system in which changes do not create new privacy risks, and there is a SAOP approved Privacy Impact Assessment. *Skip questions and complete certification.*

1b. Has an IT Compliance in Acquisitions Checklist been completed with the appropriate signatures?

☐ Yes. This is a new information system.

☐ Yes. This is an existing information system for which an amended contract is needed.

☐ No. The IT Compliance in Acquisitions Checklist is not required for the acquisition of equipment for specialized Research and Development or scientific purposes that are not a National Security System.

☒ No. This is not a new information system.

2. Is the IT system or its information used to support any activity which may raise privacy concerns?

NIST Special Publication 800-53 Revision 4, Appendix J, states "Organizations may also engage in activities that do not involve the collection and use of PII, but may nevertheless raise privacy concerns and associated risk. The privacy controls are equally applicable to those activities and can be used to analyze the privacy risk and mitigate such risk when necessary." Examples include, but are not limited to, audio recordings, video surveillance, building entry readers, and electronic purchase transactions.

☐ Yes. *(Check all that apply.)*

Activities			
Audio recordings	☐	Building entry readers	☐
Video surveillance	☐	Electronic purchase transactions	☐
Other (specify):			

☒ No.

3. Does the IT system collect, maintain, or disseminate business identifiable information (BII)?

As per DOC Privacy Policy: "For the purpose of this policy, business identifiable information consists of (a) information that is defined in the Freedom of Information Act (FOIA) as "trade secrets and commercial or financial information obtained from a person [that is] privileged or confidential." (5 U.S.C.552(b)(4)). This information is exempt from automatic release under the (b)(4) FOIA exemption. "Commercial" is not confined to records that reveal basic commercial operations" but includes any records [or information] in which the submitter has a commercial interest" and can include information submitted by a nonprofit entity, or (b) commercial or other information that, although it may not be exempt from release under FOIA, is exempt from disclosure by law (e.g., 13 U.S.C.)."

☒ Yes, the IT system collects, maintains, or disseminates BII.

☐ No, this IT system does not collect any BII.

4. Personally Identifiable Information (PII)

4a. Does the IT system collect, maintain, or disseminate PII?

As per OMB 17-12: "The term PII refers to information that can be used to distinguish or trace an individual's identity either alone or when combined with other information that is linked or linkable to a specific individual."

☒ Yes, the IT system collects, maintains, or disseminates PII about: *(Check all that apply.)*

☐ DOC employees
☐ Contractors working on behalf of DOC
☐ Other Federal Government personnel
☒ Members of the public

☐ No, this IT system does not collect any PII.

If the answer is "yes" to question 4a, please respond to the following questions.

4b. Does the IT system collect, maintain, or disseminate Social Security numbers (SSNs), including truncated form?

☐ Yes, the IT system collects, maintains, or disseminates SSNs, including truncated form.

Provide an explanation for the business need requiring the collection of SSNs, including truncated form.
Provide the legal authority which permits the collection of SSNs, including truncated form.

☒ No, the IT system does not collect, maintain, or disseminate SSNs, including truncated form.

4c. Does the IT system collect, maintain, or disseminate PII other than user ID?

☒ Yes, the IT system collects, maintains, or disseminates PII other than user ID.

☐ No, the user ID is the only PII collected, maintained, or disseminated by the IT system.

4d. Will the purpose for which the PII is collected, stored, used, processed, disclosed, or disseminated (context of use) cause the assignment of a higher PII confidentiality impact level?
Examples of context of use include, but are not limited to, law enforcement investigations, administration of benefits, contagious disease treatments, etc.

☐ Yes, the context of use will cause the assignment of a higher PII confidentiality impact level.

☒ No, the context of use will not cause the assignment of a higher PII confidentiality impact level.

If any of the answers to questions 2, 3, 4b, 4c, and/or 4d are "Yes," a Privacy Impact Assessment (PIA) must be completed for the IT system. This PTA and the SAOP approved PIA must be a part of the IT system's Assessment and Authorization Package.

CERTIFICATION

☒ The criteria implied by one or more of the questions above **apply** to the Open Data-Big Data Master System (OD-BD MS) and as a consequence of this applicability, a PIA will be performed and documented for this IT system.

☐ The criteria implied by the questions above **do not apply** to the Open Data-Big Data Master System (OD-BD MS) and as a consequence of this non-applicability, a PIA for this IT system is not necessary.

System Owner Name: Scott Beliveau Office: OCTO-EAAB Phone: (571) 272-7343 Email: Scott.Beliveau@uspto.gov Signature: ________ Digitally signed by Users, Beliveau, Scott Date: 2022.08.05 08:21:20 -04'00' Date signed: ________	**Chief Information Security Officer** Name: Don Watson Office: Office of the Chief Information Officer (OCIO) Phone: (571) 272-8130 Email: Don.Watson@uspto.gov Signature: Users, Watson, Don ________ Digitally signed by Users, Watson, Don Date: 2022.08.25 09:27:21 -04'00' Date signed: ________
Privacy Act Officer Name: Caitlin Trujillo Office: Office of General Law (O/GL) Phone: (571) 270-7834 Email: Caitlin.Trujillo@uspto.gov Signature: Caitlin D. Trujillo ________ Digitally signed by Users, Trujillo, Caitlin Date: 2022.08.16 16:38:16 -04'00' Date signed: ________	**Bureau Chief Privacy Officer and Authorizing Official** Name: Henry J. Holcombe Office: Office of the Chief Information Officer (OCIO) Phone: (571) 272-9400 Email: Jamie.Holcombe@uspto.gov Signature: Users, Holcombe, Henry ________ Digitally signed by Users, Holcombe, Henry Date: 2022.08.25 13:42:24 -04'00' Date signed: ________
Co-Authorizing Official Name: N/A Office: N/A Phone: N/A Email: N/A Signature: ________ Date signed: ________	