Architectural Blueprint for an Enterprise-Grade Conversational AI Dataset on ENSA Khouribga Institutional Knowledge

Executive Summary

This report presents a comprehensive architectural blueprint for the development of a production-grade conversational AI training dataset, with a primary focus on the institutional knowledge of the École Nationale des Sciences Appliquées de Khouribga (ENSAK). The strategic objective is to transform the institution's currently fragmented and decentralized information landscape into a coherent, accessible, and intelligent conversational experience. This document outlines a robust dataset architecture designed for scalability and maintainability, defines the core knowledge domains and intent classifications, details the feature engineering approach for precise entity extraction, and proposes a governance framework for long-term data quality. The proposed architecture addresses the challenge of a distributed digital footprint by embedding data provenance into its core schema, ensuring reliability. Furthermore, the analysis of ENSAKH's academic and research activities reveals a strategic focus on Data Science, Artificial Intelligence, and Cybersecurity, a specialization that this dataset is designed to prominently feature. The report concludes with a phased implementation roadmap and key performance indicators to guide the successful deployment of a conversational AI that can effectively serve prospective students, current attendees, faculty, and external partners.

Section 1: Foundational Architecture of the ENSAKH Conversational AI Dataset

The foundational architecture is engineered to support a scalable, maintainable, and high-performance conversational AI system. It prioritizes compatibility with modern Natural Language Understanding (NLU) engines and facilitates collaborative development among diverse stakeholders.

1.1 Dataset Schema and Format

The recommended format for the dataset is JSONL (JSON Lines). In this format, each line of a file is a self-contained, valid JSON object that represents a single training example. This approach offers significant advantages over a single large JSON array, as it is highly performant for streaming and processing large datasets, is easily version-controlled using systems like Git, and remains human-readable, which simplifies collaboration between technical teams and non-technical domain experts.

The core schema for each JSON object is defined as follows:

- intent_id: A unique, hierarchical string that identifies the user's goal (e.g., academics.initial_training.gi.get_objectives). This structure allows for granular intent classification and logical grouping.
- text_examples: An array of strings containing various user utterance examples that map to the intent. This list should capture diverse phrasing, terminology, and questions.
- entities: An array of objects defining labeled data within the text_examples. Each entity object specifies the entity's label (e.g., @program_name), its start and end character positions, and its normalized value.
- responses: An array of templated response strings. Providing multiple variations allows the AI to deliver more natural and less repetitive answers.
- metadata: An object containing critical data for provenance and quality assurance, including source_snippet_id (to trace the information back to its origin), last_validated_date, and a confidence_score.

A critical challenge identified is ENSAKH's fragmented digital footprint, with institutional knowledge distributed across at least three primary domains: the official university website (ensak.usms.ac.ma), a separate portal for continuing education (ensak-formations.com), and a network-level site (ensamaroc.com).[1] This decentralization creates a significant risk of encountering conflicting or outdated information. A user inquiring about a program should receive a single, authoritative answer, regardless of the source. Training an AI on contradictory data would directly undermine its reliability and erode user trust.
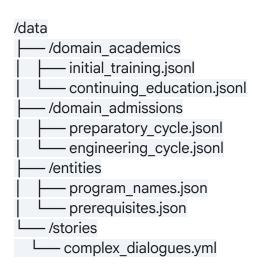
To mitigate this, the metadata field within the schema is of paramount importance. By systematically tagging every piece of information with its source_snippet_id, the dataset maintains full data provenance. A validation process, detailed in Section 5, will establish a

"source of truth" hierarchy, assigning the highest priority to information from the official ensak.usms.ac.ma domain. When data ingestion processes identify conflicting facts, the system will flag them for manual review or automatically default to the highest-priority source, thereby proactively managing the risk of misinformation.

1.2 Directory Structure and Version Control

The entire dataset will be managed within a Git-based repository to leverage robust version control, branching, and collaborative workflows. The recommended directory structure is designed to be modular, mirroring the knowledge domains outlined in Section 2. This organization enables parallel development by different domain experts, simplifies maintenance, and allows for targeted model retraining on specific domains as information evolves, without requiring a full model rebuild.

An example of the proposed directory structure is as follows:

```
/data
├── /domain_academics
│   ├── initial_training.jsonl
│   └── continuing_education.jsonl
├── /domain_admissions
│   ├── preparatory_cycle.jsonl
│   └── engineering_cycle.jsonl
├── /entities
│   ├── program_names.json
│   └── prerequisites.json
└── /stories
    └── complex_dialogues.yml
```

Section 2: Core Knowledge Domains and Intent Classification

This section deconstructs the entirety of ENSAKH's institutional knowledge into a structured taxonomy of intents. This classification serves as the semantic core of the conversational AI, defining the scope of its understanding and response capabilities.

2.1 Institutional Identity & Governance

This domain covers fundamental information about the school's identity, leadership, and contact details.

- **History & Mission:** Intents will cover the school's establishment in 2007 as part of the national initiative to train 10,000 engineers, its mission to invigorate the regional socio-economic environment, and its official integration into Sultan Moulay Slimane University on April 19, 2018.[4]
- **Governance & Administration:** Intents are designed to identify key administrative personnel, such as the Director, Pr. KADIRI Moulay Saddik, and the General Secretary, Mme JALIL Amina.[8] This sub-domain also includes the organizational structure, covering departments like the Secretariat, Human Resources, and Communication and Internships.[9]
- **General Contact & Location:** This includes intents for providing the official address (Bd Béni Amir BP N° 77 Khouribga Maroc), primary phone numbers (0523492335), and official email contacts (contact.ensak@usms.ma).[1]

2.2 Academic Offerings: Initial Training

This domain is architected around the five-year Ingénieur d'État (State Engineer) program, which is distinctly divided into a two-year integrated preparatory cycle (API1, API2) followed by a three-year engineering cycle with specialization.[3] Each engineering specialization (filière) will have a dedicated set of intents.

- **Génie Informatique (GI):** Intents will detail the program's objective of training versatile computer engineers, potential career outcomes such as Software Developer and Database Administrator, and specific admission prerequisites.[12]
- **Informatique et Ingénierie des Données (IID):** This sub-domain will cover the program's focus on data exploitation, management, analysis of large data volumes, machine learning, and its associated career paths like Data Scientist and Big Data Architect.[11]
- **Génie Électrique (GE):** Intents will focus on the program's core areas, including

embedded systems, industrial automation, process control, and microelectronics design.[12]

- **Ingénierie des Réseaux Intelligents et Cybersécurité (IRIC):** This specialization addresses the growing demand for skills in modern networking and cybersecurity. Initial data will be synthesized from comparable programs at other institutions to build a foundational dataset, which can be refined once specific ENSAKH curriculum details are obtained.[11]
- **Génie des Procédés, de l'Energie et de l'Environnement (GPEE):** This area covers the design and optimization of industrial processes with a strong emphasis on energy efficiency and environmental respect.[12]
- **Génie Réseaux et Télécommunications (GRT):** Intents for this specialization focus on mastering telecommunications networks and solving related problems.[5]
- **Management et Gouvernance des Systèmes d'Information (MGSI):** This specialization addresses the intersection of technology and business management.[11]

2.3 Academic Offerings: Advanced & Continuing Education

This domain covers post-engineering cycle programs and professional development offerings.

- **LMD System:** This includes intents related to the Licence-Master-Doctorat framework, specifically covering programs like the Master in "Informatique et Mathématiques pour la Science des Données" (IMSD) and the broader doctoral cycle, including information on the thesis charter and defense procedures.[1]
- **Formation Continue (Continuing Education):** This large sub-domain sources information primarily from the dedicated ensak-formations.com website.[2] It will feature intents for each professional Licence and Master's program offered, such as those in Artificial Intelligence, Big Data & Cloud, and Cybersecurity. Key features to be highlighted include the weekend-based course schedule and the official status of the diplomas as State University Degrees jointly issued by l'Université Sultan Moulay Slimane and ENSA de Khouribga.[2]

2.4 Admissions & Enrollment

This domain provides detailed guidance on the application and admission processes for all academic paths.

- **Preparatory Cycle (Post-Baccalaureate):** Intents will detail the national written

competitive exam (concours national écrit), the pre-selection process based on Baccalaureate grades, and the subjects tested, namely Mathematics and Physics-Chemistry.[5]

- **Engineering Cycle (Bac+2/3/4 Parallel Admission):** This covers admissions for candidates with prior higher education diplomas (e.g., DEUG, DUT, Licence, Master). Intents will describe the three-phase process: dossier review, a written exam, and a final interview. The criteria evaluated during the interview, such as communication skills, creativity, and motivation, will also be included.[3]
- **Continuing Education:** This section will detail the specific admission process for professional programs, including the list of required documents (legalized copies of CIN, Baccalaureate, and all post-Bac diplomas) and eligibility pathways, such as for candidates holding a Specialized Technician diploma from OFPPT.[2]

2.5 Research & Innovation

This domain showcases the school's research activities and practical innovations.

- **Laboratories:** Intents will describe the two primary multidisciplinary research laboratories: the "Laboratoire des Sciences et Techniques pour l'Ingénieur" (LASTI) and the "Laboratoire d'Ingénierie des Procédés, Informatique et Mathématiques" (LIPIM). Information will include their directors, research teams, and specific research axes, such as the "Mathématiques et Sciences des Données" team within LIPIM.[1]
- **Doctoral Cycle:** This includes intents covering the doctoral program (Cycle doctoral), the thesis charter, and procedures for thesis and habilitation defenses.[1]
- **GitHub Repositories:** A unique and valuable source of information is the official @ENSAK-USMS GitHub organization.[17] Intents will be created to describe popular student repositories, such as Master-Big-Data-Business-Intelligence and AlumniTrackingPlatform. This showcases the practical application of skills taught in the curriculum, offering tangible proof of the school's technical output.[17]

2.6 Campus Life & Student Services

This domain addresses the non-academic aspects of student life at ENSAKH.

- **Student Governance and Activities:** Intents will cover the central role of the Student Office (Bureau des Élèves - BDE) in representing students and organizing key events like the integration week for new students and the annual ENSA-ENTREPRISE forum.[15]

Information on student clubs and associations, such as Enactus and Jeunes Leaders Marocains (JLM), will also be included.[1]

- **Facilities & Services:** This sub-domain provides practical information about campus resources. Intents will cover the library and its online catalog, the campus-wide "Campus Connecté" WiFi network, the FabLab innovation space, and the mandatory student health insurance program (AMOE).[1] It is also important to note the absence of a university restaurant or on-campus housing, which are common queries.[13]

2.7 Partnerships & External Relations

This domain highlights the school's connections with industry and academia.

- **Industry Certifications:** A key selling point for the institution is its partnerships with leading technology companies, including Huawei, Cisco, AWS, and Fortinet. Intents will detail the availability of professional certification training in high-demand fields such as Artificial Intelligence, Cloud Computing, Internet of Things (IoT), and Cybersecurity.[22]
- **Academic Collaborations:** This includes intents describing international partnerships, such as the formal convention with INSA of Toulouse, and collaborations with Canadian universities.[1] Opportunities for double diplomas will also be covered.[1]
- **Diploma Recognition:** Intents will confirm that ENSAKH delivers a Diplôme d'ingénieur d'État (State Engineer Diploma), a nationally recognized and highly valued qualification.[3] While no specific accreditation from international bodies like the CTI was found in the available information, the dataset will be structured to easily incorporate this information if it becomes available.[24] Services used for diploma verification, such as Verifdiploma, will also be mentioned.[25]

The convergence of multiple data points—including the "Informatique et Ingénierie des Données (IID)" program, the "Master en Sciences des données et Big Data," the research focus of the LIPIM laboratory, tangible student projects on GitHub like the Master-Big-Data-Business-Intelligence repository, and industry partnerships with AWS and Huawei—indicates a deliberate strategic positioning of ENSAKH as a hub for Data Science and AI.[5] This is not a coincidental overlap but a clear institutional identity responding to national engineering needs.[5] Consequently, the dataset architecture must reflect this strategic focus. A high-level intent, such as query_institutional_strengths, will be created. Dialogue flows will be designed to synthesize information from the academics, research, and partnerships domains to provide a comprehensive answer to user queries like, "What is ENSA Khouribga known for?" This approach ensures the AI can articulate the institution's key specializations effectively.

Section 3: Feature Engineering and Entity Extraction

This section defines the structured data elements, or "entities," that the NLU model will be trained to extract from user queries. A well-defined entity dictionary is the cornerstone of a sophisticated conversational AI, allowing it to move beyond simple keyword matching to understand the specific parameters of a user's request. This enables the handling of complex, multi-faceted queries such as "Compare the admission requirements for GI and IID for a student with a DUT."

3.1 Entity Dictionary

The following table provides a comprehensive catalog of the primary entities to be recognized by the system, along with their types, descriptions, examples, and source references.

Entity Name	Type	Description & Examples	Source Snippets
@program_name	List/Synonym	The official name and acronyms of academic programs. *Ex: "Génie Informatique", "GI", "Informatique et Ingénierie des Données", "IID", "IMSD"*	[5]
@degree_level	List/Synonym	The level of study or diploma. *Ex: "Cycle Préparatoire", "Cycle Ingénieur", "Licence",*	[11]

		"Master", "Doctorat"	
@admission_req	List/Synonym	Specific diplomas or exams required for admission. *Ex: "Baccalauréat", "DEUG", "DUT", "Concours National Commun"*	3
@duration	Regex	The length of a program or cycle. *Ex: "5 ans", "3 ans", "six semestres"*	10
@department_name	List/Synonym	The academic departments. *Ex: "Mathématiques et Informatique", "Génie Electrique"*	5
@lab_name	List/Synonym	The research laboratories. *Ex: "LASTI", "LIPIM"*	1
@partner_org	List/Synonym	Names of partner universities and companies. *Ex: "INSA de Toulouse", "Huawei", "Cisco", "AWS"*	17
@contact_person	List/Synonym	Names of key administrative or academic staff. *Ex: "KADIRI Moulay Saddik", "JALIL Amina"*	8

@campus_service	List/Synonym	On-campus services and facilities. *Ex: "Bibliothèque", "FabLab", "WiFi Campus Connecté", "AMOE"*	1
@event_name	List/Synonym	Names of recurring events. *Ex: "Forum ENSA-ENTREPRISE", "semaine d'intégration"*	15

Section 4: Conversational Flows and Dialogue Management

This section outlines the design of multi-turn dialogues, or "stories," that enable the AI to handle complex user journeys. The goal is to create a helpful guide that can navigate ambiguity and gather necessary information, rather than a simple fact-retrieval bot.

4.1 Core Principles

The dialogue management system will be built upon three core principles:

- **Disambiguation:** When a user's query is ambiguous, the AI will proactively seek clarification. For instance, a query like "Tell me about the computer science program" will trigger a disambiguation prompt: *"Are you interested in Génie Informatique (GI) or the more specialized Informatique et Ingénierie des Données (IID)?"*
- **Contextual Slot Filling:** For complex, multi-parameter queries such as checking admission eligibility, the AI will guide the user by asking for necessary information sequentially. It will maintain context throughout the conversation to fill required "slots." For example: *"To check your eligibility for the engineering cycle, could you please tell me your highest completed diploma (e.g., Baccalauréat, DUT, Licence)?"*
- **Comparative Dialogue:** The AI will be explicitly designed to handle comparative queries.

Upon recognizing a request to compare two or more entities, it will fetch structured information for each and present it in a clear, comparative format.

4.2 Sample Dialogue Flow: Program Comparison

The following flow illustrates how the system would handle a comparative query, a common use case for prospective students.

- **User:** "What's the difference between GI and IID?"
- **AI Action:**
  1. The NLU model recognizes the primary intent as compare_programs.
  2. It extracts two instances of the @program_name entity: "GI" and "IID".
  3. The dialogue manager triggers a fulfillment action that fetches structured data for both programs, specifically their objectives, key curriculum topics, and primary career outcomes.
- **AI Response:** *"Génie Informatique (GI) is a generalist program focused on core software development, databases, and systems administration. It prepares you for roles like Software Developer or IT Project Manager. In contrast, Informatique et Ingénierie des Données (IID) is a specialized program focused on Data Science, Machine Learning, and Big Data technologies, leading to careers like Data Scientist or Big Data Architect. Are you more interested in general software engineering or data specialization?"*

Section 5: Data Augmentation, Validation, and Maintenance Strategy

This section defines the governance framework required to ensure the dataset's initial quality, its robustness against varied user inputs, and its currency over time. A high-quality dataset is the single most critical factor in the success of a conversational AI project.

5.1 Data Augmentation

To ensure the model is robust and can understand the many ways users might phrase a question, systematic data augmentation techniques will be employed. These include:

- **Paraphrasing:** Manually and automatically generating alternative phrasings for each training utterance.
- **Synonym Replacement:** Replacing keywords with their synonyms (e.g., "admission requirements" vs. "entry criteria").
- **Back-Translation:** Translating an utterance to another language and then back to the original to generate a syntactically different but semantically similar phrase.

5.2 Multi-Stage Validation Protocol

A rigorous, multi-stage validation protocol will be implemented to maintain the integrity and accuracy of the dataset.

- **Stage 1 (Automated Linting):** Pre-commit hooks will be integrated into the Git repository to automatically check for common errors before any new data is committed. This includes validating the JSONL format, checking for duplicate training examples, and ensuring entity annotations are correctly formatted.
- **Stage 2 (Peer Review):** All additions or modifications to the dataset must be submitted via a pull request. This process will enforce a mandatory review by at least one other team member, who will be responsible for checking factual accuracy against source documents and ensuring linguistic clarity and correctness.
- **Stage 3 (Cross-Domain Validation):** On a periodic basis, automated scripts and manual reviews will be conducted to ensure consistency across different knowledge domains. For example, the @program_name entity must be used and spelled consistently in the academics, admissions, and partnerships domains.

5.3 Long-Term Maintenance Workflow

Institutional information is dynamic; therefore, a clear workflow for keeping the dataset current is essential.

- **Quarterly Review Cycle:** A formal review cycle will be established, assigning ownership of each knowledge domain file (e.g., initial_training.jsonl) to a specific domain expert. These owners will be responsible for verifying the accuracy of the data in their respective files against official ENSAKH sources on a quarterly basis.
- **Update Process:** Any identified change—such as a new program offering, an updated admission deadline, or a change in administrative staff—will trigger the creation of a new branch and a pull request. This change will proceed through the full multi-stage validation protocol. Once merged, it will trigger an automated CI/CD pipeline to retrain

the NLU model on the updated dataset and deploy the new version.

Section 6: Implementation Blueprint and Recommendations

This final section provides a strategic, phased roadmap for building, deploying, and measuring the success of the ENSAKH conversational AI.

6.1 Phased Rollout Strategy

A phased approach is recommended to deliver value quickly while managing complexity and iteratively improving the system based on user feedback.

- **Phase 1 (Minimum Viable Product - MVP):** The initial phase will focus on the highest-value and most frequently asked questions, primarily targeting prospective students.
  - **Domains:** admissions, academics.initial (at an overview level), and institution (focusing on contact information and location).
  - **Goal:** To provide immediate, tangible value by answering the most critical questions for the largest user group, thereby reducing the burden on administrative staff.
- **Phase 2 (Expansion):** This phase will add depth and breadth to the knowledge base to serve a wider audience, including current students and faculty.
  - **Domains:** Detailed curriculum information for all academics programs, academics.advanced (Master's and PhD), partnerships (industry and academic), and campus_life.
  - **Goal:** To evolve the AI into a comprehensive resource for the entire ENSAKH community.
- **Phase 3 (Advanced Capabilities):** The final phase will incorporate complex conversational flows and specialized knowledge domains.
  - **Domains:** research (labs, publications, projects), advanced dialogue management for comparative analysis, and potential integration with external systems via APIs (e.g., live search of the library catalog).
  - **Goal:** To position the AI as a truly intelligent and indispensable institutional knowledge hub.

6.2 Key Performance Indicators (KPIs) for Success

Success will be measured through a combination of NLU performance metrics, user-facing analytics, and operational efficiency metrics.

- **NLU Performance:**
  - **Intent Recognition Accuracy:** The model's ability to correctly classify a user's intent should exceed 95% on a held-out test set.
  - **Entity Extraction F1-Score:** The F1-score for extracting critical entities like @program_name and @admission_req should be above 90%.
- **User-Facing Performance:**
  - **Query Resolution Rate:** The percentage of user conversations that are successfully concluded without needing to be escalated to a human agent. The target for this KPI is over 80%.
  - **User Satisfaction (CSAT):** Measured via a simple, non-intrusive "Was this helpful?" prompt after each response. The target is to achieve over 85% positive feedback.
- **Maintenance & Operations:**
  - **Time-to-Update:** The total time from an official information change (e.g., a new admission date published on the website) to the dataset being updated, validated, and the new model being deployed. The target for this operational metric is less than 48 hours.

Works cited

1. ENSA Khouribga – Université Sultan Moulay Slimane, accessed October 21, 2025, http://ensak.usms.ac.ma/
2. Inscription - ENSA Khouribga Formation Continue, accessed October 21, 2025, https://ensak-formations.com/inscription/
3. ENSA MAROC - Home page, accessed October 21, 2025, https://www.ensamaroc.com/
4. École nationale des sciences appliquées de Khouribga - Wikipedia, accessed October 21, 2025, https://en.wikipedia.org/wiki/%C3%89cole_nationale_des_sciences_appliqu%C3%A9es_de_Khouribga
5. École nationale des sciences appliquées de Khouribga - Wikipédia, accessed October 21, 2025, https://fr.wikipedia.org/wiki/%C3%89cole_nationale_des_sciences_appliqu%C3%A9es_de_Khouribga
6. Présentation - ENSA Khouribga, accessed October 21, 2025, http://ensak.usms.ac.ma/ensak/presentation/
7. Ecole Nationale des Sciences Appliquées- Khouribga, accessed October 21, 2025, https://www.tawjihmaroc.com/school/61
8. Mot du Directeur - ENSA Khouribga, accessed October 21, 2025,

http://ensak.usms.ac.ma/ensak/mot-du-directeur/
9. Direction – ENSA Khouribga, accessed October 21, 2025, http://ensak.usms.ac.ma/ensak/la-direction/
10. ENSA Khouribga - Infoschool -, accessed October 21, 2025, https://infoschool.ma/ensa-khouribga/
11. Formation initiale - ENSA Khouribga, accessed October 21, 2025, http://ensak.usms.ac.ma/ensak/formation-initiale/
12. Cycle Ingénieur - ENSA Khouribga, accessed October 21, 2025, http://ensak.usms.ac.ma/ensak/cycle-ingenieur/
13. ENSA KHOURIBGA - ENSA MAROC, accessed October 21, 2025, https://www.ensamaroc.com/school/13
14. ENSA Khouribga Formation Continue – Évoluez dans votre carrière avec Des Licences et Masters Spécialisés, accessed October 21, 2025, https://ensak-formations.com/
15. ENSA Khouribga - Kezakoo SUP, accessed October 21, 2025, https://www.kezakoo.com/sup/ecole/ensa-khouribga/
16. Laboratoires – ENSA Khouribga, accessed October 21, 2025, http://ensak.usms.ac.ma/ensak/laboratoires/
17. National School of Applied Science in Khouribga - GitHub, accessed October 21, 2025, https://github.com/ENSAK-USMS
18. ENSAK-USMS/Master-Big-Data-Business-Intelligence: A … - GitHub, accessed October 21, 2025, https://github.com/ENSAK-USMS/Master-Big-Data-Business-Intelligence
19. Tous les Clubs - ENSA Khouribga, accessed October 21, 2025, http://ensak.usms.ac.ma/ensak/campus/
20. Procédure de connexion au réseau WiFi Campus Connecté - ENSA Khouribga, accessed October 21, 2025, http://ensak.usms.ac.ma/ensak/procedure-de-connexion-au-reseau-wifi-campus-connecte/
21. Catalogue en ligne Bibliothèque de l'ENSA - ENSA Khouribga, accessed October 21, 2025, http://ensak.usms.ac.ma/pmb/opac_css/
22. Formations certifiantes – ENSA Khouribga, accessed October 21, 2025, http://ensak.usms.ac.ma/ensak/formations-certifiantes/
23. ENSA Khouribga - 9rayti.Com, accessed October 21, 2025, https://www.9rayti.com/ecole/ensa-khouribga
24. Espace accréditation – CTI – Commission des Titres d'Ingénieur, accessed October 21, 2025, https://www.cti-commission.fr/accreditation
25. Verify a diploma from ENSA Khouribga with Verifdiploma, accessed October 21, 2025, https://ensa-khouribga.verifdiploma.com/
26. Départements – ENSA Khouribga, accessed October 21, 2025, http://ensak.usms.ac.ma/ensak/departements/
27. Forum entreprises ENSA Khouribga le 31 mars et le 1 avril - ReKrute.com, accessed October 21, 2025, https://www.rekrute.com/en/emplois-forum-entreprises-ensa-khouribga-le-31-mars-et-le-1-avril-conseils-carriere-3178.html